# DISTORTION ANALYSIS OF ANALOG INTEGRATED CIRCUITS

# THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

#### ANALOG CIRCUITS AND SIGNAL PROCESSING

Consulting Editor: Mohammed Ismail. Ohio State University

#### Related Titles:

NEUROMORPHIC SYSTEMS ENGINEERING: Neural Networks in Silicon, edited by Tor Sver Lande; ISBN: 0-7923-8158-0

**DESIGN OF MODULATORS FOR OVERSAMPLED CONVERTERS,** Feng Wang, Rame, Harjani, ISBN: 0-7923-8063-0

SYMBOLIC ANALYSIS IN ANALOG INTEGRATED CIRCUIT DESIGN, Henrik Floberg, ISBI 0-7923-9969-2

SWITCHED-CURRENT DESIGN AND IMPLEMENTATION OF OVERSAMPLING A CONVERTERS, Nianxiong Tan, ISBN: 0-7923-9963-3

CMOS WIRELESS TRANSCEIVER DESIGN, Jan Crols, Michiel Steyaert, ISBN: 0-7923-9960 DESIGN OF LOW-VOLTAGE, LOW-POWER OPERATIONAL AMPLIFIER CELLS, R. Hogervorst, Johan H. Huijsing, ISBN: 0-7923-9781-9

VLSI-COMPATIBLE IMPLEMENTATIONS FOR ARTIFICIAL NEURAL NETWORKS, Si Mehdi Fakhraie, Kenneth Carless Smith, ISBN: 0-7923-9825-4

CHARACTERIZATION METHODS FOR SUBMICRON MOSFETs, edited by Hisham Haddai ISBN: 0-7923-9695-2

LOW-VOLTAGE LOW-POWER ANALOG INTEGRATED CIRCUITS, edited by Wouter Serdi ISBN: 0-7923-9608-1

INTEGRATED VIDEO-FREQUENCY CONTINUOUS-TIME FILTERS: High-Performan Realizations in BiCMOS, Scott D. Willingham, Ken Martin, ISBN: 0-7923-9595-6

FEED-FORWARD NEURAL NETWORKS: Vector Decomposition Analysis, Modelling and Anal Implementation, Anne-Johan Annema, ISBN: 0-7923-9567-0

FREQUENCY COMPENSATION TECHNIQUES LOW-POWER OPERATIONAL AMPLIFIERS, Ruud Easchauzier, Johan Huijsing, ISBN: 0-7923-9565-4

ANALOG SIGNAL GENERATION FOR BIST OF MIXED-SIGNAL INTEGRATE CIRCUITS, Gordon W. Roberts, Albert K. Lu, ISBN: 0-7923-9564-6

INTEGRATED FIBER-OPTIC RECEIVERS, Aaron Buchwald, Kenneth W. Martin, ISBN: 0-792 9549-2

MODELING WITH AN ANALOG HARDWARE DESCRIPTION LANGUAGE, H. Al Mantooth, Mike Fiegenbaum, ISBN: 0-7923-9516-6

LOW-VOLTAGE CMOS OPERATIONAL AMPLIFIERS: Theory, Design and Implementation Satural, Mohammed Ismail, ISBN: 0-7923-9507-7

ANALYSIS AND SYNTHESIS OF MOS TRANSLINEAR CIRCUITS, Remco J. Wiegerink, ISB 0-7923-9390-2

COMPUTER-AIDED DESIGN OF ANALOG CIRCUITS AND SYSTEMS, L. Richard Carle Ronald S. Gyurcsik, ISBN: 0-7923-9351-1

HIGH-PERFORMANCE CMOS CONTINUOUS-TIME FILTERS, José Silva-Martínez, Mich Steyaert, Willy Sansen, ISBN: 0-7923-9339-2

SYMBOLIC ANALYSIS OF ANALOG CIRCUITS: Techniques and Applications, Lawrence Huelsman, Georges G. E. Gielen, ISBN: 0-7923-9324-4

DESIGN OF LOW-VOLTAGE BIPOLAR OPERATIONAL AMPLIFIERS, M. Jeroen Fonder Johan H. Huijsing, ISBN: 0-7923-9317-1

STATISTICAL MODELING FOR COMPUTER-AIDED DESIGN OF MOS VLSI CIRCUIT Christopher Michael, Mohammed Ismail, ISBN: 0-7923-9299-X

SELECTIVE LINEAR-PHASE SWITCHED-CAPACITOR AND DIGITAL FILTERS. Husse

## DISTORTION ANALYSIS OF ANALOG INTEGRATED CIRCUITS

by

#### Piet Wambacq

IMEC, Leuven, Belgium

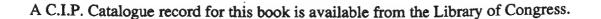
and

#### **Willy Sansen**

Katholieke Universiteit Leuven, Belgium



KLUWER ACADEMIC PUBLISHERS
BOSTON / DORDRECHT / LONDON



#### ISBN 0-7923-8186-6

Published by Kluwer Academic Publishers, P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America by Kluwer Academic Publishers, 101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed by Kluwer Academic Publishers, P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved
© 1998 Kluwer Academic Publishers, Boston
No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the copyright owner.

### **Foreword**

The analysis and prediction of nonlinear behavior in electronic circuits has long been a topic of concern for analog circuit designers. The recent explosion of interest in portable electronics such as cellular telephones, cordless telephones and other applications has served to reinforce the importance of these issues. The need now often arises to predict and optimize the distortion performance of diverse electronic circuit configurations operating in the gigahertz frequency range, where nonlinear reactive effects often dominate. However, there have historically been few sources available from which design engineers could obtain information on analysis techniques suitable for tackling these important problems.

I am sure that the analog circuit design community will thus welcome this work by Dr. Wambacq and Professor Sansen as a major contribution to the analog circuit design literature in the area of distortion analysis of electronic circuits. I am personally looking forward to having a copy readily available for reference when designing integrated circuits for communication systems.

Robert G. Meyer
Professor
Electrical Engineering and Computer Sciences
University of California
Berkeley, California
1998

### **Preface**

In the world of electronics nowadays very advanced systems can be integrated on one chip. This is mainly possible by the ability to build complex functions with digital VLSI. However, not every functionality can be achieved using digital electronics. For example, some applications might require signal processing at a high frequency that is too high to process with digital circuitry. In that case, the signals must be processed with analog circuitry which can be integrated completely or partially.

Other applications require an interface between the digital electronics and the outside world, which behaves in an analog way. As a result, such interface functions are implemented with analog electronics, which again can be integrated.

The above considerations indicate that analog integrated circuits are required, not only as integrated circuits on their own, but also as part of large mixed-signal integrated circuits. In such mixed-signal systems more and more functions are implemented in a digital way. The specifications of the circuitry that remains to be designed in the analog domain are often very tough. As a result, circuit designers not only need to concentrate on the first-order characteristics of analog circuits, which can already be very complicated and which are most often attributed to the behavior of the linearized circuit. In addition, characteristics such as nonlinear behavior may become very critical.

In addition to mixed-signal applications that demand tough specifications for the analog blocks, there are some applications where the suppression of nonlinear behavior is of utmost importance. Examples are audio applications and telecommunication applications. In the latter applications, nonlinear behavior of the circuits induces intermodulation products, which together with the noise, increase the amount of "unwanted signals", thereby lowering the performance.

In the last few years, an increased interest is seen in the integration of analog high-frequency front-ends for telecommunication applications, both in CMOS and in BiCMOS. These technologies are quickly scaling to smaller dimensions, such that they can be used at ever increasing frequencies. In this way, these silicon technologies form a cheap alternative for GaAs. CMOS technologies are cheaper than BiCMOS technologies, but the bipolar (npn) transistors of the BiCMOS technologies are in general superior at high frequencies than the MOS transistors. Integrated silicon RF front ends are found in literature for wireless communications [Long 95], used for example in the GSM standard [Seven 91, Stet 95], the DECT standard [Daw 97], and the Japanese standard for the personal handy-phone system (PHS) [Sato 96], further in GPS (Global Positioning System) receivers [Herm 91], wireless local area networks (WLAN) [Madi 96, Har 96], in applications in the ISM bands (Industrial-Scientific-Medical) around 900*MHz* [Hull 96],

2.4 GHz [Mey 97] or 5.7 GHz [Voi 97], in the North American Digital Cellular (NADC) hand-set [Kara 96], and so on.

In analog RF front-ends for communication circuits, specifications that are related to nonlinear circuit behavior are very important. For example, if a receiver front-end is not very linear, then large incoming signals from the antenna will induce much extra distortion. Such large incoming signal may be a wanted signal but it can also be a strong unwanted signal at an adjacent frequency. The increased efforts of the analog design community in the silicon integration of analog high-frequency RF front-ends has been one of the motivations to write this book.

Not only the nonlinear behavior of the RF part of a communication circuit is important to control. In many communication circuits analog integrated filters are used, for example to perform the anti-aliasing filtering function right before an analog-digital converter in a receiver. The nonlinear distortion of these filters can degrade the overall performance of a transceiver. Very often  $g_m$ -C filters are used in transceivers [Gopi 90, Chang 97], since they can achieve high speeds. This high speed is achieved at the expense of a reduced nonlinearity. In essence, a  $g_m$ -C cell is an open-loop transconductor. We shall see in this book that the conversion of a voltage into a current, which is realized by a transconductor, is difficult to realize with a high linearity without an overall feedback with a large loop gain. Hence, nonlinear distortion is an important aspect in the design of an active analog integrated filter.

In many applications of analog circuits one is only interested in the circuits' steady-state behavior in response to a sinusoidal excitation or a combination of such excitations. Indeed, many circuit aspects are easier to characterize in the steady state. This is partially due to the fact that an extremely large class of analog circuits can be approximated very well by a linear system. Since sinusoidal functions are eigenfunctions of linear systems, the latter ones can be easily characterized in terms of responses to sinusoidal excitations. Examples of quantities that characterize a circuit in steady state are transfer characteristics like gain or impedances. These characteristics are also best measured when a circuit is in steady state. Gain and impedances are mainly due to the behavior of the linearized circuit. However, most analog integrated circuits behave weakly nonlinearly. This means that, when a sinusoidal signal or a combination of sinusoids is applied to a circuit, the output spectrum does not only contain signals with the same frequency of the input signals, as one would expect of a linear system: in addition, the output spectrum contains small components — usually unwanted — at frequencies other than the input signal frequencies. If one sinusoidal signal is applied at the input, then these unwanted signals occur at multiples of one of the input frequencies. In this case they are termed as harmonics. In the case of an excitation with more than one sinusoidal signal, unwanted components occur at frequencies which are linear combinations of the input frequencies. These components are denoted as intermodulation products.

The weakly nonlinear behavior of analog integrated circuits is caused by the slight curvature of the characteristics of the devices of the circuit around the operating point. This behavior contrasts with *strongly nonlinear* behavior, where devices such as transistors switch between an on-state and an off-state. Nonlinear behavior, both weakly and strongly nonlinear, are not always unwanted. For example, oscillators and mixers explicitly rely on nonlinearities for a suitable operation. This book is concentrated on weakly nonlinear behavior only. In this way, the majority of continuous-time analog integrated building blocks is covered.

The harmonics or intermodulation products characterize the amount of nonlinearity of a given circuit. Since sinusoidal signals and sums of sinusoids are frequently used as inputs, a sinusoidal steady-state analysis of weakly nonlinear circuits is certainly not restrictive. Such analysis is carried out in the frequency domain.

The familiarity of circuit engineers with linear systems has given rise to many useful insights and design rules for circuits that can be approximated as linear. A circuit engineer is able to derive closed-form expressions for characteristics of a linearized circuit, which he can interpret and use afterwards during the synthesis of the circuit. If the circuit or its simplified schematic is too large to analyze with hand calculations, he can resort to a symbolic analyzer for linear(ized) circuits, but he still remains able to some extent to reason about the linear circuit's behavior even without explicitly having expressions for the characteristics.

On the other hand, the analysis and synthesis of circuits in which nonlinearities play a role, is difficult. Indeed, for such circuits just a few design rules exist in the analog design world. This has several reasons First, circuit designers are trained to reason only about linear systems, but not about nonlinear ones. Secondly, it is not easy to analyze nonlinear effects by hand calculations. The most studious designers use Taylor series, but this approach is only feasible in small circuits at low frequencies, with a very small number of nonlinearities. Usually, nonlinear effects are analyzed with tedious time-domain simulations (so-called transient simulations), followed by a Fourier analysis. Other approaches such as the harmonic balance techniques, although very useful, are not (yet) universally used. With both approaches the simulation results are numbers. These numbers can be plotted onto a graph, which can give valuable information, but they do not indicate the fundamental circuit parameters that determine the observed performance. As a result, circuit designers often do not know in which way a circuit can be improved in order to meet the specifications related to nonlinear behavior.

The above problem could be relieved if it were possible to indicate to circuit designers which circuit elements or which effects are mainly responsible for the observed nonlinear behavior. Such insight will be offered in this book for several building blocks of analog integrated circuits. In addition some general concepts of nonlinear behavior of analog integrated circuits will be studied as well.

This book is intended as a guide to learn designers of analog integrated circuits to reason about nonlinear phenomena in weakly nonlinear, continuous-time analog ICs. The required background to read this book is an understanding of analog integrated circuit design. A prior knowledge about theory of nonlinear systems, such as the theory of Volterra series, is not required. The background of Volterra series that is required to understand some essential concepts, is contained in this book.

When browsing through this book, the reader will notice the large number of formulas in this book. When one wants to reason about nonlinear phenomena, then a minimum amount of mathematics is required. It is virtually impossible to write a comprehensible text on nonlinear effects without mathematics. However, no special advanced mathematical techniques are used in this book. Also, the mathematics are explained as clearly as possible, they are interpreted and illustrated with examples, and tedious derivatives are moved to appendices at the end of the book.

This book is devoted to the analysis in the frequency domain of weakly nonlinear circuits. The circuits that are addressed are building blocks of analog integrated circuits, both in bipolar

and CMOS technologies. The emphasis is on getting insight, both in the nonlinearities in the transistors themselves, as in the nonlinear behavior of transistor circuits.

The outline of the text is as follows:

- Chapter 1 presents an overview of the approach that is followed in this book. Further, some assumptions are made about the nonlinear circuits that will be analyzed in this book, and the scope of the book will be outlined.
- In the analog design community, many definitions and keywords are used to characterize the nonlinear behavior of analog circuits in the frequency domain. This basic terminology is described in Chapter 2.
- In order to analyze the nonlinear behavior of a circuit, we need to describe the different nonlinearities in the circuit under consideration with a sufficient accuracy. To this purpose, we will make use of power series expansions of the model equations that describe a nonlinear device. It will turn out that a device such as a transistor consists of several basic nonlinear elements, such as nonlinear conductances, transconductances and capacitors. Each of these elements can be described with a power series. The coefficients of these power series are proportional to the derivatives of the model equations that describe these basic nonlinear elements. These power series coefficients, further in the text referred to as *nonlinearity coefficients* determine the nonlinear behavior behavior of a circuit. In Chapter 3, these power series are presented together with some simple examples.
- A very useful technique to describe the nonlinear behavior of weakly nonlinear circuits is the Volterra series approach, which is covered in Chapter 4. With Volterra series, it is possible to take into account frequency effects into the calculations of harmonics and intermodulation products. This is absolutely necessary if one wants to study circuits with capacitors, both linear and nonlinear. Further, we will use Volterra series to study some general concepts in nonlinear circuits: the use of feedback, both linear and nonlinear, the exploitation of symmetry for the suppression of even-order or odd-order harmonics, the effect of cascading nonlinear circuits, and pre- and post-distortion will be studied.
- Calculation methods for harmonics and intermodulation products are discussed in Chapter 5. The emphasis is on methods that allow to generate closed-form expressions for harmonics or intermodulation products. Numerical methods will be discussed briefly.

For the generation of closed-form expressions a calculation method can be used that makes use of Volterra series. This method is explained with an example. Derivations can be found in literature [Buss 74, Chua 79b]. Further, an alternative method is developed in this chapter that yields the same results without making use of Volterra series.

With both methods the circuit is analyzed in the frequency domain. In fact, we perform an AC analysis on a circuit in which the nonlinear elements are not only represented by their linearized equivalent: in addition, the second- and third-order nonlinear behavior are taken into account as well.

Despite the availability of different methods to obtain closed-form expressions for harmonics or intermodulation products, the use of these methods for hand calculations is too tedious, even for very small circuits. With a symbolic network analysis program, however, it is possible to automate these hand calculations and to obtain a closed-form expression. Symbolic network analysis programs can compute a closed-form expression for the AC behavior of a linearized circuit as a function of the symbolic small-signal parameters of the circuit and of the complex frequency variable. Modern symbolic analysis programs, such as the program ISAAC [Giel 89, Giel 91], can even generate approximate symbolic expressions. The approximation is made because the exact expression is usually too lengthy, such that it cannot be easily interpreted. An approximation based upon numerical values for the circuit parameters can retain the few dominant terms of an expression, such that the resulting expression becomes interpretable. In Chapter 5 an extension of the program ISAAC is described for the generation of approximate, interpretable expressions for nonlinear distortion.

- The nonlinearity coefficients of the different nonlinearities in a bipolar transistor and a MOS transistor are discussed in Chapters 6 and 7, respectively. Whereas for the bipolar transistor the Gummel-Poon model is still widely used or it forms the basis of more recent models [Mc And 95], the situation for a MOS transistor is more complicated. Due to the rapid scaling of a MOS transistor, effects that were recognized as second-order effects in older technologies, become dominant in modern devices. If these effects are not included in a transistor model or if they are badly modeled, then this may lead to very large errors on the harmonics or intermodulation products. The reason is that the nonlinearity coefficients are proportional to higher-order derivatives. The error on these derivatives tends to increase dramatically when the model equation is inaccurate. In Chapter 7 some shortcomings of widely used transistor models will be discussed. Further, a model for the drain current will be presented that will be used to derive closed-form expressions for the nonlinearity coefficients.
- Apart from the examples that have been used throughout the individual chapters, some more applications are described in Chapter 8. Distortion will be analyzed for the following circuits: a single-transistor amplifier, both a bipolar and a MOS version, a bipolar and a MOS differential pair, a source follower, an emitter follower, a bipolar transistor with emitter degeneration, a common-base bipolar and a common-gate MOS transistor, a bipolar and a MOS current mirror, a Miller-compensated operational amplifier, a bipolar double-balanced mixer, and a CMOS upconverter. Hereby, use will be made of the extension of the program ISAAC to generate closed-form expressions for distortion.
- Instead of computing the nonlinearity coefficients, it is also possible to measure these coefficients. A measurement procedure and measurement results on bipolar transistors are given in Chapter 9.

Finally the authors would like to thank the following persons for their discussions and comments: Peter Kinget from Lucent Technologies, Murray Hill, Yannis Tsividis from the University of Columbia, New York, Stéphane Donnay, Hugo De Man and Luc Dupas from IMEC,

Leuven, Georges Gielen, Walter Daems and Joos Vandewalle from the ESAT Laboratories of the Catholic University of Leuven, Petr Dobrovolny from the University of Brno, Czech Republic, Guang-Ming Yin and Frederic Stubbe from Rockwell Semiconductor Systems, Newport Beach, California, Frank Op 't Eynde from Alcatel Microelectronics, Brussels, Paco Fernández from IMSE-CNM, Seville, Spain, Jan Vanthienen, and, last but not least, Kaat François for her support and patience.

Piet Wambacq Willy Sansen

TO OUR FAMILIES Kaat, Lien, and Elli Hadewych, Katrien, Sara, and Marjan

### List of symbols and abbreviations

 $C_{\mu}$ 

 $C_{\pi}$ 

The SPICE model parameters are not included.	They can be found in [Hs	pi 96, Anto 88].
--	--------------------------	------------------

The SPICE model parameters are not inclu-	ded. They can be found in [Hspi 96, Anto 88].
	multiplication symbol in symbolic expressions. Normally omitted, only used for clarity.
×	multiplication symbol in numbers, for example $1.5 \times 10^{-11}$ .
$a^*$	the complex conjugate of the complex number $a$ .
$eta_F$	transistor "beta": maximum value of ratio between the collector current and the base current of a bipolar transistor.
$\gamma$	body-effect coefficient of a MOS transistor.
$arepsilon_{ox}$	$3.4531 \times 10^{-11} F/m$ , dielectric permittivity of $SiO_2$ .
$arepsilon_{Si}$	$1.0359 \times 10^{-10} F/m$ , dielectric permittivity of $SiO_2$ .
$\lambda$	channel-length modulation factor (MOS transistor).
$\mu$	surface mobility (MOS transistor).
$\phi$	surface inversion potential (MOS transistor).
heta	mobility-reduction coefficient (MOS transistor).
arg(z)	phase of the complex number $z$ .
BJT	bipolar junction transistor.
C	symbol for a capacitor.

tor).

base-collector capacitance (bipolar transis-

base-emitter capacitance (bipolar transistor).

$C_{cs}$	collector-substrate capacitance (bipolar transistor).
$C_{db}$	bulk-drain capacitance (MOS transistor).
$C_{gb}$	gate-bulk capacitance (MOS transistor).
$C_{gd}$	gate-drain capacitance (MOS transistor).
$C_{gs}$	gate-source capacitance (MOS transistor).
$C_{sb}$	bulk-source capacitance (MOS transistor).
$C_{ox}^{\prime}$	MOS gate oxide capacitance per unit area.
$C_{ox}$	total MOS gate oxide capacitance.
CMRR	common-mode rejection ratio.
$\det(s)$	determinant of the admittance matrix of a line ear network as a function of the complex frequency variable $s$ .
$E_c$	critical electric field (MOS transistor).
$E_{\it eff}$	average normal electric field that is experienced by carriers in the inversion layer (MOS transistor).
$f'(x), \frac{df}{dx}$	two notations for the derivative of $f$ with respect to $x$ .
g,G	symbols that are used for conductances.
GBW	gain-bandwidth product.
$g_m$	transistor transconductance (MOS and bipe lar transistor).
$g_{mb}$	bulk transconductance (MOS transistor).
$g_o$	transistor output conductance (MOS at bipolar transistor).
$g_\pi$	incremental base-emitter conductance (bip lar transistor).
$\mathbf{H}_n$	nth-order Volterra operator.
$HD_2$ , $HD_3$	second- and third-order harmonic distortion
$H_n(j\omega_1,j\omega_2,\ldots,j\omega_n)$	n-dimensional Fourier transform of the nt order Volterra kernel, also denoted as nt order transfer function.
$h_n( au_1, au_2,\ldots, au_n)$	nth-order Volterra kernel (time-domain representation).

 $i_{OUT}$  $I_{OUT}$  $i_{out}$  $I_{out}$  $i_B, i_b, I_B, I_b$  $i_C, i_c, I_C, I_c$  $i_D, i_d, I_D, I_d$  $i_{DSAT}, i_{dsat}, I_{DSAT}, I_{dsat}$  $I_{KF}$  $IM_2$ ,  $IM_3$ IMFDR $IP_{2h}$ ,  $IP_{3h}$  $IP_{2i}$ ,  $IP_{3i}$  $I_S$  $I_{SE}$ jk

total current through a component (time domain). DC component of the current through a component. AC component of the current through a component (time domain). Hence  $i_{OUT} = I_{OUT} +$ iout. phasor of the current through a component in the steady state (sinusoidal excitation). total value (time domain), AC value, DC value and phasor respectively of the base current of a bipolar transistor. total value (time domain), AC value, DC value and phasor respectively of the collector current of a bipolar transistor. total value (time domain), AC value, DC value and phasor respectively of the drain current of a MOS transistor in the triode region. total value (time domain), AC value, DC value and phasor respectively of the drain current of a MOS transistor in saturation. forward knee current (bipolar transistor). second- and third-order intermodulation distortion. intermodulation-free dynamic range. second- and third-order intercept point for harmonics. second- and third-order intercept point for intermodulation products. saturation current (bipolar transistor).

base-emitter leakage saturation current (bipo-

 $1.38062 \times 10^{-23} J/K$ , Boltzmann's constant.

lar transistor).

 $\sqrt{-1}$ 

 $K_{nx}$ 

 $K'_{nr}$ 

 $K_{m_{jg_1\&(m-j)g_2}}$ 

 $K_{m_{jg_1}\&kg_2\&(m-j-k)g_3}$ 

nth-order nonlinearity coefficient in the power series expansion of the function that describes the nonlinear relationship between the current and the controlling voltage for either a nonlinear conductance, transconductance or capacitor. The symbol x represent the linearized equivalent of the nonlinear element.

normalized nonlinearity coefficient:  $K_{n_x}$  divided by x.

mth-order nonlinearity coefficient in the two dimensional power series expansion of th function that describes the nonlinear relation ship between the current and the control ling voltages u and v for a two-dimensional transconductance. The symbols  $g_1$  and  $g_2$  represent the coefficients in the power series of the first-order terms in u and v, respectively. If the nonlinear relationship is expressed at i = f(u, v), then  $K_{m_{jg_1}\&(m-j)g_2}$  is given by:

$$\frac{\partial^m f(u,v)}{\partial u^j \partial v^{m-j}} \cdot \frac{1}{j!} \cdot \frac{1}{(m-j)!}$$

If j or (m-j) are equal to one, then they are usually omitted, like in  $K_{2q_1 \& q_2}$ .

mth-order nonlinearity coefficient in the three-dimensional power series expansion of the function that describes the nonlinear relationship between the current and the controlling voltages u,v and w for a three dimensional transconductance. The symbol  $g_1$ ,  $g_2$  and  $g_3$  represent the coefficients in the power series of the first-order terms in u, and w, respectively. If the nonlinear relationship is expressed as i = f(u, v, w), the  $K_{m_{jg_1}\&kg_2(m-j-k)g_3}$  is given by:

$$\frac{\partial^m f(u,v,w)}{\partial u^j \partial v^k \partial w^{m-j-k}} \cdot \frac{1}{j!} \cdot \frac{1}{k!} \frac{1}{(m-j-k)!}$$

Note that, if j, k or (m - j - k) are equation one, then they are usually omitted, like  $K_3$ <sub> $q_1$ </sub>& $q_2$ & $g_3$ .

effective channel length of a MOS transistor

 $N_A$  $n_E$  $n_F$  $n_i$  $P_{-1dR}, P_{-3dR}$  $Q_B$  $Q_{B0}$  $Q_I'(x)$ r, R $R_B, r_B$  $R_{Bex}, r_{Bex}$  $R_{Bi}, r_{Bi}$ **RF**  $r_o$  $r_{\pi}$ sT $TF_{i_1 \rightarrow output}$ 

acceptor concentration in the bulk region of a MOS transistor. base-emitter emission coefficient (bipolar transistor). forward emission coefficient (bipolar transistor). carrier intrinsic concentration  $(1.45\times10^{10}cm^{-3})$  at room temperature). 1dB or 3dB compression point. 1.6022 C, elementary charge. majority charge in the neutral base region (bipolar transistor). majority charge in the neutral base region at  $v_{BC} = 0V$ . inversion layer charge per unit area at the position x in the channel of a MOS transistor  $(0 \le x \le L)$ . symbols that are used for resistances; r = 1/gand R = 1/G. DC and AC base resistance (bipolar transistor). DC and AC extrinsic base resistance (bipolar transistor). In SPICE this is denoted by RBM.DC and AC intrinsic base resistance (bipolar transistor). radio frequency. transistor output resistance (bipolar and MOS

transistor output resistance (bipolar and MOS transistor).

incremental base-emitter resistance (bipolar transistor).

complex frequency variable (Laplace transform variable).

absolute temperature (in degrees Kelvin). transfer function from a current source  $i_1$  to the output of interest, which can be a node

voltage or a current through a circuit element.

THD	total harmonic distortion.
$t_{ox}$	gate-oxide thickness (MOS transistor).
$V_{AF}$	(forward) Early voltage (bipolar transistor).
$v_{CONTR}$	total value of a voltage that controls a nonlinear circuit element (time domain).
$V_{CONTR}$	DC component of the voltage that controls nonlinear circuit element.
$v_{contr}$	AC component (time domain) of the voltation that controls a nonlinear circuit element (time domain). Hence $v_{CONTR} = V_{CONTR} + v_{cont}$
$V_{contr,m.n}$	phasor of the component at the frequence $ m\omega_1+n\omega_2 $ of the voltage that controls a not linear circuit element (two sinusoidal excit tions with frequencies $\omega_1$ and $\omega_2$ ).
$v_{oldsymbol{pq}}$	AC component (time domain) of the difference between the voltage at node $p$ and $q$ $v_{pq} = v_p - v_q$ .
$V_{pq,m,n}$	phasor of the component at the frequence $ m\omega_1 + n\omega_2 $ of the difference between the voltage at node $p$ and $q$ : $V_{pq,m,n} = V_{p,m,n} - V_{q,m,n}$ .
$v_{DSAT}$	drain-source saturation voltage (MOS transistor).
$v_{sat}$	thermal velocity of carriers, also denoted a saturation velocity.
$V_{TO}$	zero-bias gate-source extrapolated threshold voltage of a MOS transistor.
$V_T$	gate-source extrapolated threshold voltage of a MOS transistor.
$V_t$	the thermal voltage $kT/q$ (25.86 $mV$ at root temperature).
W	effective channel width of a MOS transistor.
$\omega, \omega_1, \omega_2$	pulsation (= 2 $\pi$ * frequency) $(rad/sec)$ .

### **Contents**

	Foreword					
	Prefa	reface				
List of symbols and abbreviations						
	Cont	ents			xvii	
1	Intro	duction			1	
	1.1	Scope of	of this book	k	. 4	
2	Basic	c termir	ology		7	
_	2.1				. 7	
	2.2			excitation		
	2.3			dd-order distortion only		ļ
	2.4			ccitation		
	2.5		1			ŀ
	2.6	Cross r	nodulation		. 24	
	2.7	Summa	ary		. 25	j
3	Desc	ription	of nonline	earities in analog integrated circuits	26	,
	3.1	_			. 26	)
	3.2	Power	series desc	cription of basic nonlinearities	. 29	)
		3.2.1		r conductance and transconductance		
			3.2.1.1	Example: collector current of a BJT	. 32	)
			3.2.1.2	Example: base current of a BJT		5
		3.2.2	Nonlinear	r resistance	. 34	ŀ
			3.2.2.1	Conversion formulas between a voltage-controlled description		
				and a current-controlled description	. 35	,
			3.2.2.2	Difference between DC and AC resistance	. 37	7
		3.2.3	Nonlinea	r capacitance	. 38	3
			3.2.3.1	Example: diffusion capacitance		)
		3.2.4	Two-dim	ensional transconductance	. 39	)

			3.2.4.1 Example: two-dimensional collector current	40
		3.2.5	Three-dimensional transconductance	41
			3.2.5.1 Example: drain current of a MOS transistor	42
		3.2.6	Tracking nonlinearities	45
	3.3	Integra	ted resistors	47
		3.3.1	Nonlinearity coefficients of an implanted or diffused resistor	48
			3.3.1.1 Example	51
	3.4	Integra	ted capacitors	52
	3.5	_	y nonlinear transistor models: introduction	55
	3.6	Summa	ary	57
4			ies and their applications to analog integrated circuit design	59
	4.1		action	59
	4.2		of Volterra series	61
		4.2.1	Volterra operators	62
		4.2.2	Time-domain fundamentals	63
		4.2.3	Frequency-domain representation	65
		4.2.4	Weakly nonlinear circuit behavior revisited	67
	4.3		bles of Volterra kernels	68
		4.3.1	Basic second-order system	68
		4.3.2	Basic third-order system	69
		4.3.3	Application: a nonlinear amplifier	71
	4.4		near performance parameters in terms of Volterra kernels	75
		4.4.1	Single-tone and two-tone definitions	75
		4.4.2	Cross modulation	79
	4.5		ession of even-order or odd-order kernels	
		4.5.1	Application: suppression in a differential pair	
		4.5.2	Application: suppression in a Gilbert multiplier	
	4.6		de connection of nonlinear systems	
		4.6.1	General expressions	
		4.6.2	Application: a two-stage amplifier	88
	4.7		stortion and post-distortion using inverse systems	89
		4.7.1	General expressions	
		4.7.2	Example	
		4.7.3	Applications	
	4.8		and nonlinear feedback	- 3
		4.8.1	Nonlinear feedback systems	
		4.8.2	Feedback with a large loop gain	
		4.8.3	Linear feedback	- 4
			4.8.3.1 Simplification to memoryless systems	
		4.0.4	4.8.3.2 Application: emitter degeneration	107
		4.8.4	Nonlinearities in the feedback network	103
		4 X >	LOAGING ETTECT OF A HORISINEAL TEEGDACK HELWOLK	IVJ

CONTENTS	xix
OO - 1	

		4.8.6	Operational amplifier in a linear feedback configuration	05		
		4.8.7	Nonlinear feedback applications			
	4.9		ary			
5	Calc	culation of harmonics and intermodulation products				
J	5.1		iction	16		
	5.2		ation of Volterra kernels			
	J.2	5.2.1	First-order kernels			
		5.2.2	Second-order kernels			
		5.2.3	Third-order kernels			
		5.2.4	Postprocessing of the results	28		
		5.2.5	Simplifications			
		5.2.6	Volterra kernels of currents			
		5.2.7	Interpretation of the results	33		
		5.2.8	Factorization of the denominators			
	5.3	Direct	calculation of nonlinear responses	37		
		5.3.1	First-order responses	39		
		5.3.2	Second-order responses	143		
		5.3.3	Third-order and higher-order responses	47		
		5.3.4	Interpretation and factorization	155		
	5.4	Symbo	olic computation of harmonics and intermodulation products			
		5.4.1	Symbolic network analysis of linearized analog circuits			
		5.4.2	Symbolic analysis of weakly nonlinear analog circuits with ISAAC 1			
			5.4.2.1 Elimination of unimportant nonlinearities			
			5.4.2.2 Generation of the approximate symbolic subexpressions 1			
	5.5	-	e example circuits			
		5.5.1	Nonlinear resistive voltage divider			
		5.5.2	Nonlinear capacitive current divider			
	5.6		rical verification with other methods			
		5.6.1	Numerical integration			
		5.6.2	Shooting methods			
		5.6.3	Harmonic balance methods			
	<i>-</i> -	5.6.4	Example: an emitter follower			
	5.7	Summ	ary	178		
6	Sili	con bipo	olar transistor models for distortion analysis	180		
	6.1	Introdu	uction	180		
	6.2	The co	ollector current	182		
		6.2.1	Collector current with a linear Early effect	182		
		6.2.2	Nonlinearity of the Early effect			
	6.3		ase current			
(1)	6.4		ase resistance			
		6.4.1	Modeling of the current dependence	192		

		6.4.2	DC and AC base resistance and the nonlinearity coefficients	. 193
		6.4.3	Evaluation of the nonlinearity coefficients	
	6.5	Capaci	itors in a bipolar transistor	
	6.6	Summ	ary	. 19 <b>9</b>
7	MO	S transi	istor models for distortion analysis	200
	7.1	Introdu	uction	. 200
	7.2	Nonlin	nearity coefficients of the three-dimensional drain current nonlinearity	. 203
		7.2.1	Coefficients referred to the source	. 206
		7.2.2	Coefficients referred to the bulk	. 206
		7.2.3	Relationship between the coefficients of the two reference systems	. 208
	7.3		relations for the drain current in strong inversion	
	7.4	Drain	current in the triode region without small-geometry effects	. 212
		7.4.1	Uniform depletion layer	
			7.4.1.1 Application: a single-transistor mixer	
			7.4.1.2 Formulation of the current in terms of $v_{GB}$ , $v_{DB}$ and $v_{SB}$	
		7.4.2	Nonuniform depletion layer	
		7.4.3	Simplification: linearly varying depletion layer	
		7.4.4	Comparison of nonlinearity coefficients	
	7.5		current in saturation without small-geometry effects	
		7.5.1	Uniform depletion layer	
		7.5.2	Nonuniform depletion layer	
		7.5.3	Simplification: linearly varying depletion layer	
		7.5.4	Comparison of nonlinearity coefficients	
	7.6		ive mobility	
		7.6.1	Mobility model of Sabnis and Clemens	
		7.6.2	Drain current in the triode region	
		7.6.3	Drain current in the saturation region	
		7.6.4	Other mobility models	
			7.6.4.1 The mobility model of Frohman-Bentchkowsky	
			7.6.4.2 The mobility model of Liang et al	
		7.6.5	Evaluation of nonlinearity coefficients	
			7.6.5.1 Triode region	
		T7 1 '	7.6.5.2 Saturation region	
	7.7		ty saturation	
		7.7.1	Velocity-field models	
		7.7.2	Drain current in the triode region	
			7.7.2.1 Drain current with the simple velocity-field models	
			7.7.2.2 The functions large, mobred and hot	
			7.7.2.3 Merging of the functions <i>mobred</i> and <i>hot</i>	
			7.7.2.4 Drain current with the more accurate velocity-field model	
		772	7.7.2.5 Modelling around $v_{DS} = 0V$	
		7.7.3	Drain current in the saturation region	. 200

			7.7.3.1	Drain current in saturation with the simple velocity-field models	253
			7.7.3.2	Merging of the functions mobred and hot	255
			7.7.3.3	Drain current in saturation with the more accurate velocity-	
				field model	256
		7.7.4	Evaluation	on of nonlinearity coefficients	257
			7.7.4.1	Nonlinearity coefficients for the triode region	258
			7.7.4.2	Approximate expressions for the nonlinearity coefficients in	
				the triode region	258
			7.7.4.3	Nonlinearity coefficients for the saturation region	267
			7.7.4.4	Approximate expressions for the nonlinearity coefficients in	
				the saturation region	268
	7.8	Nonun	iform dop	ing effects	273
		7.8.1	Modelin	g with one single body-effect coefficient	276
		7.8.2	Adaption	of the threshold voltage expression	276
		7.8.3		rrent model with three equations	
		7.8.4		reduction	
	7.9	Thresh	old voltag	e for short- and narrow-channel devices	278
	7.10	Source	and drain	resistances	280
		7.10.1	Source a	nd drain resistance components in conventional devices	281
		7.10.2	LDD str	uctures	281
				the drain current and on the nonlinearity coefficients	
	7.11	The ou	tput cond	uctance and its derivatives in saturation	283
		7.11.1	The phy	sical model of Huang et al	284
			7.11.1.1	Contribution of channel-length modulation	286
			7.11.1.2	Contribution of drain-induced barrier lowering	287
			7.11.1.3	Contribution of the substrate current	288
			7.11.1.4	Continuity of the output conductance	289
			7.11.1.5	Evaluation of the output conductance and its derivatives	289
			7.11.1.6	Evaluation of other nonlinearity coefficients in saturation	. 292
	7.12	Capac	itors in a l	MOS transistor	. 292
		7.12.1	Extrinsi	c capacitors	. 293
				capacitors	
	7.13	Drain	current in	weak inversion operation	. 297
		7.13.1	Express	ion of the drain current	. 297
		7.13.2	Nonline	arity coefficients of the drain current in weak inversion	. 298
	7.14	Summ	ary		. 300
8	Was	kly nor	ilinear bo	havior of basic analog building blocks	302
•	8.1	-			
	8.2			ransistor amplifier	
	0.∠	8.2.1	-	tary transistor model	
		0.2.1	8.2.1.1	Computation of harmonics from the DC transfer characteristic	
			8.2.1.2	Computation of harmonics with the method of Section 5.3	
			··		,

xxii CONTENTS

	8.2.2	Influence of the output resistance	. 310
		8.2.2.1 Fundamental response	. 310
		8.2.2.2 Second harmonic	
		8.2.2.3 Third harmonic	. 314
	8.2.3	Influence of the source resistance	. 315
		8.2.3.1 Fundamental response	
		8.2.3.2 Second harmonic	
		8.2.3.3 Third harmonic	
	8.2.4	Influence of the base resistance	
		8.2.4.1 Fundamental response	. 320
		8.2.4.2 Second harmonic	
		8.2.4.3 Third harmonic	. 323
	8.2.5	Influence of $C_{\pi}$	. 325
		8.2.5.1 Current drive	
		8.2.5.2 Voltage drive	
	8.2.6	Influence of $C_{\mu}$ and $C_{cs}$	
		8.2.6.1 Fundamental response	
		8.2.6.2 Second harmonic distortion	
		8.2.6.3 Third harmonic distortion	. 343
		8.2.6.4 Third harmonic distortion with different values of $C_{\mu}$	. 347
8.3	Single	MOS transistor amplifier	. 350
	8.3.1	Influence of $g_m$ only	. 350
		8.3.1.1 Transistor in the saturation region	. 352
		8.3.1.2 Transistor in the triode region	. 354
		8.3.1.3 Transistor in the weak inversion region	
	8.3.2	Influence of the output conductance	. 355
	8.3.3	Frequency behavior	. 357
		8.3.3.1 First-order response	
		8.3.3.2 Second harmonic distortion	. 360
		8.3.3.3 Third harmonic distortion	. 365
8.4	Bipola	ar differential pair	. 366
	8.4.1	Computation of harmonics from the DC transfer characteristic	
	8.4.2	Symbolic analysis of $HD_3$	. 369
	8.4.3	$HD_2$ due to mismatches	. 372
8.5	MOS	differential pair	. 375
	8.5.1	Computation of harmonics from the DC transfer characteristic	
	8.5.2	Computation of $HD_3$ including the bulk effect	
	8.5.3	$HD_2$ due to mismatches	
8.6	Emitte	er follower	
	8.6.1	First-order response	. 387
	8.6.2	Second-order response	. 388
	8.6.3	Third-order response	
8.7	Source	e follower	. 391

CO	NTE	NTS	xiii
		8.7.1 First-order response	392
		8.7.2 Second-order response	392
		8.7.3 Third-order response	395
	8.8	Cascode transistor	
		8.8.1 Bipolar cascode	
		8.8.1.1 First-order response	398
		8.8.1.2 Second-order response	399
		8.8.1.3 Third-order response	401
		8.8.2 MOS cascode	401
	8.9	Common-gate and common-base transistor	401
		8.9.1 First-order response	403
		8.9.2 Second-order response	404
		8.9.3 Third-order response	406
	8.10	Current mirrors	407
		8.10.1 DC transfer characteristic for a bipolar current mirror	407
		8.10.2 Distortion in a MOS current mirror	409
		8.10.2.1 First-order response	410
		8.10.2.2 Second-order response	410
		8.10.2.3 Third-order response	412
		8.10.2.4 Numerical example	412
		8.10.3 Distortion in a bipolar current mirror	413
	8.11		414
	8.12	CMOS Miller-compensated operational amplifier	417
	8.13	CMOS upconverter	420
	8.14	Summary	427
9	Mea	asurements of basic nonlinearities of transistors	429
	9.1	Introduction	429
	9.2	Principle of the measurements	430
	9.3	Practical applications	431
	9.4	Measurement results	433
		9.4.1 Derivatives of $i_C$ with respect to $v_{BE}$	434
		9.4.2 The nonlinearity of the Early resistance	436
		9.4.3 Cross-derivatives of $i_C$	438
	9.5	Summary	439
	Bib	liography	440
	Ap	pendices	456
A	A Useful trigonometric relationships		

В	Basics of Volterra series 458			
	B.1	Introduction		
	B.2	Volterra series representation of a system		
	B.3	Second-order Volterra systems		459
		B.3.1 The sec	cond-order operator	460
		B.3.2 The sec	cond-order Volterra operator	460
		B.3.3 Second	l-order kernel symmetrization	463
	<b>B</b> .4			
			o-dimensional Fourier and Laplace transform	
		B.4.2 Sinuso	idal response of a second-order Volterra system	464
			nse of a second-order system to a sum of two sinusoids	
	B.5	Higher-order Volterra systems		
			h-order operator	
			h-order Volterra operator	
			ler kernel symmetrization	
			dimensional Laplace and Fourier transforms	
C	Derivation of the method for the direct computation of nonlinear responses 476			
	C.1	Setup of basic equations		
	C.2	First-order responses		
	C.3			
	C.4			
D	Nonlinearity coefficients for the description of the Early effect			478
E	Rela	Relation between source-referred and bulk-referred nonlinearity coefficients of a		
	MOS transistor			482
F	Deri	vatives of the d	rain current with an implicit saturation voltage	486
	F.1	Determination	of $v_{DSAT}$	487
	F.2	First-order deri	ivatives	487
	F.3	Higher-order derivatives		
G	Deri	vation of the M	OS drain current in the presence of velocity saturation	490
	G.1	G.1 Derivation of the drain current with the simple velocity-field models		
	G.2	Derivation of the	he drain current with the more accurate velocity-field model	492
		G.2.1 The rig	gorous approach	492
		G.2.2 Approx	ximate approach	493
	Inde	X		495

## Chapter 1

### Introduction

In this book we will present the reader some insight into the nonlinear behavior of the basic building blocks of analog integrated circuits.

The study of the nonlinear behavior of basic building blocks will lead to insights that are applicable to larger circuits as well. In addition, we will study several general concepts of nonlinear circuits.

In order to study nonlinear behavior we need some analytical basis that will enable us to generate expressions for the nonlinear behavior, from which we can obtain some insight. The starting point for such basis will be the theory of Volterra series [Sche 80], although we will use a different technique as well, that closely resembles the Volterra series approach.

Volterra series can also be used to study general concepts of (weakly) nonlinear circuits. Volterra series describe nonlinear systems in a way that is similar to the approximation of an analytical function with a Taylor series: small excursions around an operating point can be described with very few terms of a Volterra series. The larger the excursions, the more terms of this series need to be taken into account for an accurate description. When the excursions are too large, then the series diverges. This corresponds to a signal that extends the region of convergence of the series. When we limit the signal amplitudes within the region of convergence of the Volterra series, then we can use this technique to analyze nonlinear circuits. Roughly speaking, this limitation corresponds to weakly nonlinear behavior. The circuits that are analyzed in this book are assumed to be driven by signals that are small enough such that the Volterra series converges.

In addition we assume that the nonlinear effects caused by circuits that behave in a weakly nonlinear way, is caused by second- and third-order nonlinear behavior only. In this way, we make simplifications, but instead we will be able, as will be shown in this book, to identify dominant nonlinearities. Roughly speaking, accuracy is exchanged for interpretability and insight. The fact that dominant nonlinearities can be identified, is due to the fact that with the Volterra series approach the nonlinear response can be seen as a sum of contributions from the different nonlinearities in the circuit. These contributions can be visualized and the dominant ones can be identified. In addition, the nonlinear responses are computed with the Volterra series approach by repeatedly solving sets of linear equations.

With this in mind, it is possible to obtain closed-form expressions for the nonlinear behav-

ior. These closed-form expressions only take into account the dominant nonlinearities. For circuits of practical size, these expressions cannot be easily computed by hand. Instead, one can resort to symbolic network analysis programs. These programs, which have matured in the last decade, are able to automatically generate approximate closed-form expressions of the AC behavior of linearized circuits. Since with the Volterra series an approximation of the nonlinear behavior can be obtained by repeatedly solving linear equations, existing symbolic network analysis programs can be extended to compute approximate, interpretable symbolic expressions of a circuit's nonlinear behavior. Such extension has been made with the symbolic simulator ISAAC [Giel 89, Giel 91] and is described in detail in [Wamb 96]. This extension is also discussed briefly in this book. However, more attention is paid in this book to the results obtained with this approach. Compact, interpretable expressions for nonlinear behavior that are virtually impossible to obtain by hand, can now be obtained in seconds or minutes. In this way, the designer can spend his time in the interpretation of the expressions, rather than on concentrating to the generation of the expressions.

In addition to their use for generating interpretable data for nonlinear circuit behavior, Volterra series can also be very useful to study nonlinear circuits in general. For example, the use of feedback, both linear and nonlinear, the exploitation of symmetry for the suppression of even-order or odd-order harmonics, can be studied with Volterra series. Such concepts are studied in this book for general circuits, after which they are applied to some specific circuits. In weakly nonlinear circuits, the distortion levels are small. As a result, a careful modeling of the devices will be required. Whereas the behavior of a linearized circuit is a function of the small-signal parameters of the devices, which are nothing else but first-order derivatives of the model equations of the devices, it will be seen in this book that the harmonics and the intermodulation products are determined by higher-order derivatives of the model equations. Since errors on model parameters and inaccuracies in the model equations themselves tend to increase with the order of the derivatives — computing derivatives is a numerically unstable process — it is clear once again that a careful modeling of the devices is very important.

Just as circuit designers need to reason about the behavior of a linearized circuit in terms of small-signal parameters, one will have to reason about distortion in terms of those higher-order derivatives. These derivatives are difficult to compute accurately. Accurate values can be obtained for example by device simulations. Alternatively, the higher-order derivatives can also be measured in some situations, as will be shown in this book. However, data obtained either with device simulations or with accurate measurements do not indicate which physical effects are behind those data. However, this information is very valuable, since it can yield physical insight. This information can be obtained when an expression is available for a given derivative. However, expressions for higher-order derivatives can be very complicated, especially when complicated models are used. Fortunately, the lengthy expressions for the higher-order derivatives often contain a few terms that are dominant over the rest of the terms. When these dominant terms can be identified, one can obtain insight in the physical effects that determine a derivative. This will be shown in this book.

The values for the higher-order derivatives, either supplied by device simulations, measurements, or by computations from model equations, can be used in the Volterra series approach to obtain interpretable data for harmonics or intermodulation products. This approach will be

used in order to obtain insight in the nonlinear operation of some fundamental analog building blocks, that are part of complete analog integrated functions, such as filters, transceiver frontends, .... In this book, only the nonlinear behavior of those basic building blocks will be studied. Distortion of larger circuits, consisting of several building blocks, can be used in specialized papers [Groen 94, Tsiv 94].

Since the Volterra series approach only generates approximate values for the harmonics or intermodulation products, as already mentioned above, it is useful to compare the numerical data obtained with this approach to other techniques that are more accurate while they offer less insight. Some numerical simulation methods, such as numerical integration, the harmonic balance technique and the shooting technique, are briefly discussed in this book.

**Example** We now illustrate the need to control the nonlinear behavior of an analog RF frontend. Figure 1.1 depicts a simplified architecture of a receiver front-end. It has to be noticed that several alternative architectures are possible. The front-end depicted here is only chosen for illustration purposes.

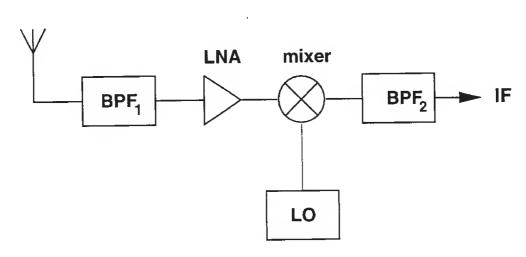


Figure 1.1: A part of a receiver front-end.

The antenna receives a complete spectrum of signals, among which the wanted signal. The goal of a receiver front-end is to select a wanted signal from the complete spectrum of signals that enter the antenna. Simply stated, this is obtained by rejecting the unwanted signals using filtering, in combination with an amplification of the wanted signal. A first bandpass filter, right after the antenna, already blocks a large part of the unwanted spectrum. The center frequency of this filter is in the middle of the frequency band that has been attributed to the application under consideration. After this bandpass filter, the signals are fed into a low-noise amplifier. Then the signals are downconverted to an intermediate frequency  $f_{IF}$ , which may be zero, by mixing those signals with a local oscillator signal at a frequency  $f_{LO}$ . This frequency is controlled by a frequency synthesizer. The value of  $f_{LO}$  is chosen such that the difference with the frequency of the wanted RF signal, denoted as  $f_{RF}$ , is equal to  $f_{IF}$  that is usually sufficiently lower than  $f_{RF}$  and  $f_{LO}$ . After the first mixer stage a bandpass filter is placed with a center frequency of  $f_{IF}$ . If

 $f_{IF} \neq 0$  then the wanted signal, which is a spectrum around  $f_{IF}$  is further downconverted (no shown on the figure).

A mixer is also sensitive for unwanted signals at the frequency  $f_{RF} - 2f_{IF}$ : these signal are also downconverted by the local oscillator to the frequency  $f_{IF}$ . In order to get rid to some extent of the signals at the frequency  $f_{RF} - 2f_{IF}$ , which is denoted as the *image frequency* another bandpass filter is placed between the low-noise amplifier and the mixer. This filter often denoted as an image-reject filter. When the RF signal is immediately downconverted to the baseband, then this filter is not needed. Instead, however, other problems arise, which are discussed for example in [Crols 95b, Abid 95].

Assume now that the front-end of Figure 1.1 is a part of a receiver for mobile telephony. In the GSM standard for mobile telephony the wanted signal is a channel with a 200kHz band width. The different channels are situated in the frequency band between 935MHz and 960MH for mobile reception and in the band from 890MHz to 915MHz for mobile transmission. The spacing between adjacent channels is 200kHz. It is possible that the power of the wanted signal is several orders of magnitude smaller than the power of adjacent signals. These adjacent signal are not rejected by one of the two filters shown in Figure 1.1, and hence they are processed by the low-noise amplifier and the mixer in much the same way as the wanted signal. These two block are not perfectly linear. It will be explained in this book that due to the nonlinear behavior of these blocks, two strong unwanted signals that are adjacent to the wanted signal, give rise to an intermodulation product that can have the same frequency as the wanted signal. This intermodulation product can be larger than the wanted signal. It is clear that in this situation the detection of the wanted signal by the rest of the front-end and the baseband circuitry will be hampered by the presence of this intermodulation product.

#### 1.1 Scope of this book

The distortion that is studied in this book is harmonic or intermodulation distortion that is caused by the weakly nonlinear behavior of the circuit devices. This type of distortion is generally referred to as nonlinear distortion. Other types of distortion exist as well. In fact, distortion is nothing else but a deviation of the output signal from a wanted waveform.

Distortion can also arise in a linear circuit. Then we speak about *linear distortion*. This kind of distortion is illustrated in the next example. It is not further considered.

Example The occurrence of linear distortion is illustrated with the opamp circuit of Figure 1.2 This circuit acts as an inverting amplifier with a voltage gain  $A_v$  of  $-R_2/R_1$ . The frequency response of this amplifier is shown in Figure 1.3. The frequency response of the operational amplifier has been represented by the transfer function  $A_v(f)$  with a dominant pole at the frequency  $f_0$ . The gain-bandwidth product of the operational amplifier is given by  $GBW = A_v(0)f_0$ . As a result, the frequency response of the overall amplifier configuration has a low-frequency value of  $-R_2/R_1$  and a pole at a frequency  $f_1$  that is a factor  $|R_1/R_2|$  lower than the gain-bandwidth product GBW of the operational amplifier.

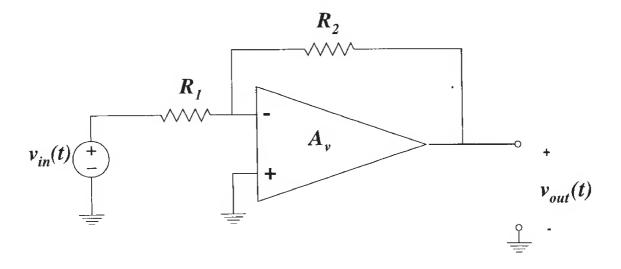


Figure 1.2: An inverting amplifier.

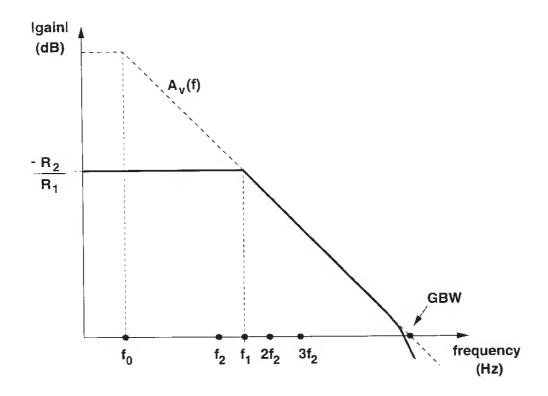


Figure 1.3: Absolute value of the frequency response of the inverting amplifier.

Assume now that a square wave signal is applied with a frequency  $f_2$  that is somewhat smaller than  $f_1$ . The response of the amplifier to this waveform is **distorted** as shown in Figure 1.4: the output waveform clearly deviates from a square wave. We assumed all circuit elements to be linear such that the distortion cannot be caused by nonlinear behavior. The observed distortion can be explained as follows. From the theory of Fourier series we know that a periodic waveform can be considered as a sum of sine waves at the harmonic frequencies of  $f_1$ . Since the frequency response of the inverting amplifier of Figure 1.2 is not flat, the different harmonics are not amplified in the same way. As a result, the sum of the amplified harmonics does no longer correspond

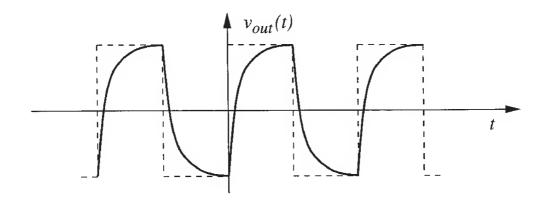


Figure 1.4: Response of the amplifier of Figure 1.2 to a square wave at frequency  $f_2$  (see Figure 1.3).

to a square wave.

Strongly nonlinear behavior is not addressed in this book neither. In general, it is not possible to obtain closed-form expressions of circuits that behave in a strongly nonlinear way. Fortunately the vast majority of analog integrated circuits behave in a weakly nonlinear way with the signal levels for which these circuits have been designed.

The exclusion of strongly nonlinear behavior implies that distortion induced by slew rate in not considered in this book. The slew rate of an amplifier indicates the maximum change of the output voltage per time unit. When a sinusoidal signal is applied to a circuit with a given slew rate, then the frequency of this signal may be so high that the amplifier cannot follow the fast changes of the signal, and the amplifier starts to slew, which means that the output signal will change at a constant speed. In this way, the response of an amplifier to a sinusoidal signal will resemble to a triangle signal [Solo 74]. This is only a simplified representation of an effect that can give rise to complicated waveforms in practical amplifiers.

Another type of this distortion that is not considered in this book is crossover distortion. This type of distortion occurs for example in class B output stages [Gray 93].

### Chapter 2

### **Basic terminology**

#### 2.1 Introduction

For many analog integrated circuits the performance with respect to their nonlinear behavior is often expressed in terms of parameters that are measured in the frequency domain. These parameters are defined in this chapter.

The definitions in this chapter are explained by looking at the output of a general nonlinear analog circuit. This circuit is excited with one or more sources of the form  $A_i \cos(\omega_i t)$   $(i=1,2,\ldots)$ . This circuit may be any continuous-time analog circuit that ideally operates in a linear way. For circuits like a mixer, which are essentially nonlinear, some extra definitions are provided. In this introductory chapter the definitions will be clarified without taking into account frequency. In Chapter 4 it will be explained how these definitions can be reformulated in terms of Volterra series. This will allow to take into account frequency effects.

In this book we will analyze responses of circuits that are excited by one or two sinusoidal signals. In the next sections it will be explained that due to the nonlinear behavior of the circuit, the output response will contain components at the frequencies  $m\omega_1 + n\omega_2$ . The amplitude of response at this frequency will be written as follows:

amplitude of response at 
$$m\omega_1 + n\omega_2 = |V_{k,m,n}|$$
 (2.1)

where  $\omega_1$  and  $\omega_2$  are the (radial) frequencies of the sinusoidal excitations and m and n are integers. The phasor of the component of a node voltage, say at node k, at the frequency  $m\omega_1 + n\omega_2$  is denoted in this book by  $V_{k,m,n}$ . This is a complex number.

### 2.2 Single-frequency excitation

We will first describe qualitatively the frequency spectrum at the output of the test circuit when it is excited with one sinusoidal source at a frequency  $\omega_1$ . When the amplitude  $A_1$  of the input signal is small enough, then the output spectrum of the circuit only contains one frequency component above the noise floor, namely the response corresponding to the circuit's linear behavior.

This is a signal at the same frequency of the input signal, called the *fundamental frequency*. The amplitude of this signal changes proportionally with the input amplitude.

When the input amplitude is increased, the output spectrum contains signals at the frequencies  $2\omega_1$  and  $3\omega_1$ . These signals, called the **second and third harmonics**, originate from **second** and third-order nonlinear circuit behavior,

respectively, as we shall see below. Harmonics higher than the third, caused by higher-order behavior, come above the noise floor at even higher input amplitudes. It is seen that the amplitude of the nth harmonic increases with the nth power of the input amplitude: an increase of the input amplitude with 6dB yields an increase of the second harmonic at the output with 12dB, the third harmonic increases with 18dB and so on. At higher input amplitudes this is not true anymore. Then it is observed that the third-order nonlinear behavior also gives rise to a component at the fundamental frequency which increases with the third power of the input amplitude. As a result the fundamental response can increase faster than linear, which for an amplifier means that the gain slightly increases. In this case one speaks about gain expansion. If, on the other hand, the increase is less than linear because the sign of the third-order contribution is opposite to the sign of the linear response, then a gain compression is observed at the output. Similarly, fourth-order behavior gives a contribution to the second harmonic and so on. This situation is depicted in Figure 2.1. Signals caused by nonlinear behavior of order higher than five are not shown in this figure. Also it must be noted that a component at 0Hz is found at the output. As we shall compute below, this DC shift is caused by second-order, fourth-order, or in general, even-order nonlinear behavior.

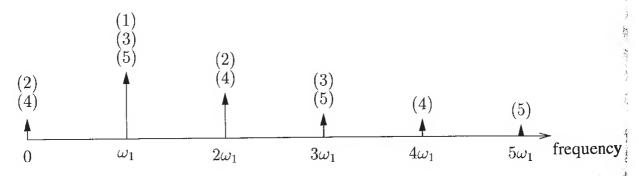


Figure 2.1: The different harmonics at the output of an analog circuit excited by a sinusoidal signal at frequency  $\omega_1$ . The numbers between brackets indicate the order of nonlinear behavior by which the signal is determined.

When the amplitude is increased even more, then the signal level in the circuit cannot increase anymore since the signal swing is limited by the power supplies or by cutoff or saturation of some active element in the circuit. In this situation the signal levels at the different harmonics do not increase much anymore.

The linear response is usually the wanted response while a harmonic is unwanted. Therefore, it is common to denote the harmonics as distortion, nonlinear distortion or harmonic distortion. The second-order harmonic distortion or briefly the second harmonic distortion  $HD_2$  and the third harmonic distortion  $HD_3$  are defined as the ratio of the second and third harmonic,  $V_{out,2,0}$ 

and  $V_{out,3,0}$  respectively, to the fundamental response  $V_{out,1,0}$ :

$$HD_2 = \frac{|V_{out,2,0}|}{|V_{out,1,0}|} \tag{2.2}$$

$$HD_3 = \frac{|V_{out,3,0}|}{|V_{out,1,0}|} \tag{2.3}$$

For higher order, of course, the definitions are similar. Also, the extension to an output current instead of a voltage is straightforward. The harmonic distortion figures are dimensionless. They are often expressed in percent or in dB for a given input amplitude. The harmonic distortion figures are often used to characterize the general nonlinear behavior of the circuit even if this circuit is not excited by one single tone. The reason is that  $HD_2$  and  $HD_3$  can be computed and measured rather easily compared to other quantitative measures that are described below.

The nonlinear behavior as discussed above in a qualitative way can be clarified mathematically with a simple example. Assume that the relationship between the input signal x(t) (which is either a current or a voltage) of a circuit and the output y(t) is given by the following relationship

$$y(t) = K_1 \cdot x(t) + K_2 \cdot (x(t))^2 + K_3 \cdot (x(t))^3 + \dots$$
 (2.4)

When the input-output relationship is given explicitly by an analytic relationship

$$y(t) = f(x(t)) (2.5)$$

then the coefficients  $K_1, K_2, K_3, \ldots$  can be identified with the coefficients of a Taylor series<sup>1</sup> of f:

$$K_1 = \frac{df}{dx} \tag{2.6}$$

$$K_2 = \frac{1}{2} \cdot \frac{d^2 f}{dx^2} \tag{2.7}$$

$$K_3 = \frac{1}{6} \cdot \frac{d^3 f}{dx^3} \tag{2.8}$$

The coefficient  $K_1$  describes the behavior of the linearized circuit. This behavior is often referred to as *first-order behavior* and the coefficient  $K_1$ . The coefficients  $K_2, K_3, \ldots$  are called **second-order** and **third-order nonlinearity coefficients of the circuit**, respectively or, in general, **higher-order nonlinearity coefficients**. The coefficients  $K_2$  and  $K_3$  determine the second- and third-order nonlinear circuit behavior.

A computation of the higher-order coefficients  $K_2$  and  $K_3$  is not straightforward. It is one of the goals of this book to explain how these coefficients can be computed. In this section we consider these coefficients as known figures. Further, it is assumed in this section for simplicity that there is no delay between the output and the input. In other words, the circuit under consideration has no memory: it contains no capacitors or inductors or the frequencies of interest are so low

It is assumed that the series of equation (2.4) converges to f(x(t)).

that capacitors and inductors do not play a role yet. Such circuits are referred to as **memoryless** circuits. Circuits in which capacitors and/or inductors play a role are denoted as circuits with **memory**. For memoryless circuits the coefficients  $K_1, K_2, K_3, \ldots$  are independent of frequency. In this book we will compute the response of nonlinear circuits both with and without memory.

Returning to our simple example, we assume that the input signal x(t) has the form

$$x(t) = A_1 \cos(\omega_1 t + \alpha_1) \tag{2.9}$$

Substituting this expression into equation (2.4) yields the output y(t):

$$y(t) = A_1 K_1 \cos(\omega_1 t + \alpha_1) + A_1^2 K_2 \left(\frac{1}{2} + \frac{1}{2}\cos(2\omega_1 t + 2\alpha_1)\right) + A_1^3 K_3 \left(\frac{3}{4}\cos(\omega_1 t + \alpha_1) + \frac{1}{4}\cos(3\omega_1 t + 3\alpha_1)\right)$$
(2.10)

Some useful trigonometric relationships to compute such responses are given in Appendix A.

It is seen that the second-order coefficient  $K_2$  gives rise to a signal at  $2\omega_1$  and at 0Hz. Both signals are proportional to  $K_2$  and to the square of the input amplitude  $A_1$ . Therefore, these signals are denoted as **second-order signals**. The third-order coefficient  $K_3$  gives rise to a signal at the frequency  $3\omega_1$  and at the fundamental frequency  $\omega_1$ . These signals, which are proportional to  $K_3$  and to the third power of the input amplitude, are referred to as **third-order signals**.

Assume that  $K_3$  has the same sign as  $K_1$ . In that case the third-order signal at the fundamental frequency has the same sign as the first-order signal. In other words, the amplitude of the fundamental signal has increased due to third-order behavior. This situation corresponds to expansion. If  $K_1$  and  $K_3$  have an opposite sign then we have compression.

Consider now the phase of the different harmonics. Compare the situation in which the phase  $\alpha_1$  in equation (2.9) is zero, to the situation with  $\alpha_1$  is 180 degrees. In the second situation the input signal is the opposite of the input signal with  $\alpha_1 = 0$ . From equation (2.10) it can be seen that a change of the sign of the input signal will change the sign of the fundamental and the third harmonic, but not of the second harmonic. This is exploited in a balanced circuit, to which two input signals of opposite phase but equal amplitude are applied. A differential output signal in a perfectly balanced circuit does not contain a second harmonic. This is also true for the other even-order harmonics. In Section 2.3 this issue is further explained.

In equation (2.10) we have restricted our calculations to order three. However, the calculations can be easily extended in order to include behavior of order higher than three. Then one would observe that fourth-order behavior will give rise to a signal at the fourth harmonic, at the second harmonic and at 0Hz. Fifth-order behavior will give rise to a signal at  $5\omega_1$ ,  $3\omega_1$  and  $\omega_1$ , and so on. This corresponds to the spectrum shown in Figure 2.1.

The circuits that we analyze in this book are nonlinear circuits that are excited by input signals with a fairly small amplitude. In this way, we can assume that the response at a harmonic frequency  $n\omega_1$  is only determined by nth-order nonlinear behavior. We will clarify this with the following example. From equation (2.10) we recall that the response at the fundamental frequency is determined not only by first-order behavior, but also by third-order and higher

odd-order behavior. Throughout this book, we will assume that the third-order signal at the fundamental frequency that is proportional to  $A_1^3K_3$  is much smaller than the first-order signal that is proportional to  $A_1K_1$ . Under this assumption the phasor of the response at the frequency  $n\omega_1$  can be written as

$$V_{out,n,0} = p_n A_1^n K_n (2.11)$$

in which  $p_n$  is a rational number. The response at  $n\omega_1$  is then given by

response at 
$$n\omega_1 = |V_{out,n,0}| \cos(n\omega_1 t + n\alpha_1)$$
 (2.12)

Further it is assumed that the input signals are so small that the energy of the output signal is mainly concentrated in the lowest-order harmonics. In this way, the amplitude of the second harmonic is much higher than the amplitude of the fourth harmonic, which in turn is much higher than the amplitude of the sixth harmonic, and so on. In other words, the array of amplitudes of the even-order harmonics is decreasing. Similarly, the array of amplitudes of the odd-order harmonics is decreasing as well. These assumptions correspond to weakly nonlinear behavior. Sometimes these assumptions or conditions are also referred to as low-distortion conditions. In Chapter 4 the notion of weakly nonlinear behavior will be defined more rigorously in terms of Volterra series. A circuit that is excited by a signal that is sufficiently small such that the above

conditions are satisfied, is often referred to as a weakly nonlinear circuit. Strictly speaking, weakly nonlinear behavior does not only depend on the circuit but also on the amplitude of the input signal. Nevertheless, the notion of weakly nonlinear circuit is often used for circuits that

operate in a weakly nonlinear way at the signal levels for which it has been designed.

It is seen that in this case the nth harmonic is proportional to the nth power of the input amplitude.

Weakly nonlinear behavior is the counterpart of *strongly nonlinear behavior*. Whereas weakly nonlinear behavior is typically due to the curvature of the characteristics of a transistor in a given operating region, strongly nonlinear behavior corresponds to the switching of a transistor between an "on-state" (e.g. the saturation region for a bipolar transistor) and an "off-state" (cut-off). Strongly nonlinear behavior occurs for example in the output stage of an amplifier that is excited by a too large signal. Other examples of strongly nonlinear behavior are slew-rate behavior [Solo 74], the behavior of a comparator and class-AB operation of an amplifier [Lak 94].

Returning now to our example, we now compute the second and third harmonic distortion figures under the assumption that the circuit behaves in a weakly nonlinear way. From the response of the circuit, given in equation (2.10) we find, using the definitions of equations (2.2) and (2.3),

$$HD_2 = \frac{1}{2}A_1 \left| \frac{K_2}{K_1} \right| \tag{2.13}$$

$$HD_3 = \frac{1}{4}A_1^2 \left| \frac{K_3}{K_1} \right| \tag{2.14}$$

In earlier years [Nar 67, Poon 72, Chis 73, Kuo 73, Nar 73], harmonic distortion was often defined in terms of a given output power. The input amplitude is then chosen such that the fundamental waveform at the output dissipates a certain power, for example  $1 \, mW$ , in a load resistance Although this concept might be appropriate for communication systems, the above definitions are more general and certainly more appropriate for circuits loaded with a high impedance which is possibly reactive.

Whereas Figure 2.1 shows the different harmonics for a given amplitude level, Figure 2.2 depicts the first three harmonics at the output as a function of the input amplitude level.

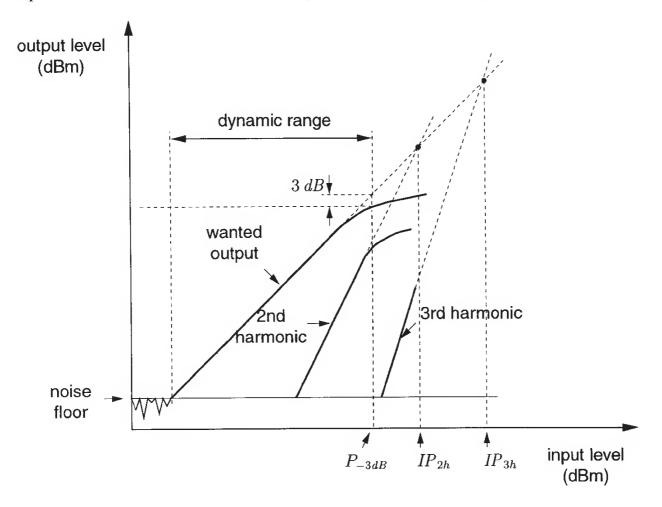


Figure 2.2: Level of the harmonics at the output of a nonlinear circuit as a function of the input amplitude.

The horizontal and vertical scales in this figure are logarithmic. In communication circuits it is common to express signal (voltage) levels in dBm. With these units the signal levels are expressed in terms of power dissipated in a  $50\Omega$  resistor. A level of 0dBm corresponds to a dissipation of 1mW in a  $50\Omega$  resistor. This is equivalent to a sinusoidal voltage with an effective value of  $\sqrt{10^{-3} \times 50} = 223.6 mV$ , or to an amplitude of 316mV. A power dissipation of P Watter then yields a value of  $10\log_{10}(P/10^{-3})dBm$ .

Another useful "logarithmic" unit for voltage levels is dBVrms. The value of a voltage in dBVrms is found by taking 20 times the logarithm with base ten of the RMS value of the signal

Compression point Returning to Figure 2.2, it is seen that from a certain input level, the fundamental response does not increase proportionally with the input level anymore. Instead, the gain has decreased a little bit. A quantitative measure for this gain compression is the *IdB* or *3dB* compression point, which are denoted as  $P_{-1dB}$  and  $P_{-3dB}$ , respectively. These figures indicate the input level for which the fundamental response is 1 or 3dB lower than the extrapolation of the response at lower amplitudes which is caused by linear behavior only.

The second harmonic slightly above the noise floor in Figure 2.2 is due to second-order non-linear behavior and hence increases with the square of the input amplitude. At higher input levels the second harmonic is also influenced by higher-order behavior. This can give compression or expansion of the second harmonic, depending on the sign of the second-order term and that of the higher-order terms that influence the value of the second harmonic. In Figure 2.2 it is seen that the second harmonic exhibits compression.

Intercept points When the input amplitude is sufficiently small, our test circuit under consideration behaves in a weakly nonlinear way. Then the fundamental response increases linearly with the input amplitude and the amplitude of the second harmonic at the output increases with the square of the input amplitude. Since the second harmonic due to second-order behavior increases faster with the input level than the fundamental response due to linear behavior, the extrapolation of both curves must intersect. The input level that corresponds to this intersection is called the **second-order intercept point**  $IP_{2h}$ . The subscript h is used here to denote "harmonic", in order to distinguish with the intercept point defined for intermodulation distortion, as will be discussed in Section 2.4. It is easy to see that this level corresponds to the input amplitude for which  $HD_2$  according to equation (2.13) is equal to one. The higher this intercept point, the better of course the second harmonic is suppressed. Similarly, a **third-order intercept point**  $IP_{3h}$  is defined. Sometimes a distinction is made between the intercept point at the input, which is the same as the intercept point as defined above, and the output intercept point. The latter is given by the output level that corresponds to the input intercept point.

Under low-distortion conditions and at low frequencies the intercept points  $IP_{2h}$  and  $IP_{3h}$  can be computed as a function of the coefficients  $K_1$ ,  $K_2$  and  $K_3$  from the general input-output relationship given in equation (2.4). Since  $IP_{2h}$  and  $IP_{3h}$  are defined as the input amplitudes at which  $HD_2$  and  $HD_3$  respectively are equal to one, we find, using equation (2.13) and (2.14),

$$IP_{2h} = 2 \left| \frac{K_1}{K_2} \right| (2.15)$$

$$IP_{3h} = 2\sqrt{\left|\frac{K_1}{K_3}\right|} (2.16)$$

Dynamic range Figure 2.2 indicates yet another parameter, namely the *input dynamic range* or, briefly, dynamic range. This is the range of the input amplitude range over which the circuit can be used without a too large signal degradation. The lower limit is determined by the input referred noise while the upper limit is usually taken equal to the 1dB or 3dB compression

point. The (input) dynamic range is the ratio of those two limits, usually expressed in dB. The corresponding range of the output signals is called the output dynamic range.

Total harmonic distortion Many applications require amplifiers with a high linearity. Examples of such applications are high resolution data acquisition systems and high-fidelity audio equipment. For such amplifiers it is important that a sinusoidal input signal yields a sinusoidal output signal of high spectral purity. In this case the higher harmonics all together represent an unwanted distortion signal. To this purpose, the *total harmonic distortion* is used to indicate how closely the output waveform resembles to a pure sine wave. This can be measured by the amount of energy in the harmonics relative to the energy in the fundamental. In this way, the total harmonic distortion *THD* is defined as the sum of the RMS values of the higher harmonics, relative to the fundamental component:

$$THD = \sqrt{\frac{\sum_{n=2}^{\infty} |V_{out,n,0}|^2}{|V_{out,1,0}|^2}}$$
 (2.17)

which, under low-distortion conditions, reduces to

$$THD = \sqrt{\sum_{n=2}^{\infty} \left(\frac{A_1^{n-1} K_n}{K_1}\right)^2}$$
 (2.18)

The value of the total harmonic distortion is often specified in dB or in %. Equation (2.18) reveals that under low-distortion conditions the total harmonic distortion increases as the input amplitude increases.

Whereas total harmonic distortion is very often not the most relevant parameter to characterize the nonlinear behavior of a circuit in the application for which the circuit is designed, the total harmonic distortion of circuits is very often specified. This is then considered as a measure of "how nonlinear a circuit is at a given input amplitude" and is often used to compare different circuits for a given application.

**Examples of** *THD* The general definition of total harmonic distortion given in equation (2.17) is illustrated with the computation of *THD* for the square wave shown in Figure 2.3. This signal could correspond to the output of a circuit that is driven by a large sinusoidal input signal such that the output stage switches between the positive and the negative power supply. This corresponds to strongly nonlinear way. In order to compute *THD*, we start from the Fourier series of this square-wave signal:

$$f(t) = \frac{4}{\pi} \left( \sin(t) + \frac{1}{3} \sin(3t) + \frac{1}{5} \sin(5t) + \dots \right)$$
 (2.19)

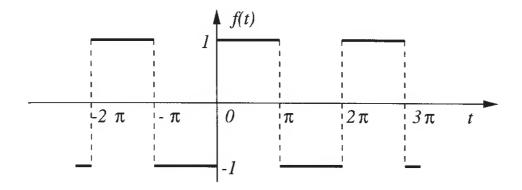


Figure 2.3: A square-wave signal.

The amplitudes of the different harmonics in this equation can be substituted in the definition of THD from equation (2.17). This yields

$$THD = \sqrt{\sum_{n=1}^{\infty} \frac{1}{(2n+1)^2}} = \sqrt{\frac{\pi^2}{8} - 1} \approx 48.3\% = -6.3dB$$
 (2.20)

It is seen that in this example THD is independent of the input amplitude. It is a very high value compared to the total harmonic distortion of so-called low-distortion circuits. For example, in [Smit 94] an operational amplifier is reported that produces harmonic distortion that is 95dB lower than the wanted signal, when the amplifier drives a  $100\Omega$  load and the input voltage has a peak-to-peak value of 1V.

As another example, we compute *THD* for the triangular waveform of Figure 2.4.

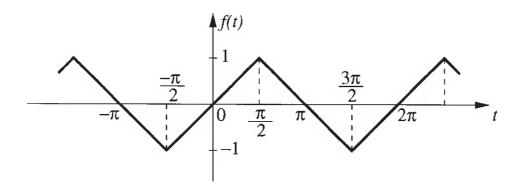


Figure 2.4: A triangular waveform.

The Fourier series of the triangular waveform is given by

$$f(t) = \frac{8}{\pi^2} \left( \sin(t) - \frac{1}{3^2} \sin(3t) + \frac{1}{5^2} \sin(5t) - \dots \right)$$
 (2.21)

Substituting the amplitudes of the different harmonics into equation (2.17) yields

$$THD = \sqrt{\sum_{n=1}^{\infty} \frac{1}{(2n+1)^4}} = \sqrt{\frac{\pi^4}{96} - 1} \approx 12.1\% = -18.3dB$$
 (2.2)

which is seen to be about 12dB (about a factor four) smaller than the total harmonic distortion of a square-wave signal.

## 2.3 Even-order and odd-order distortion only

It has already been mentioned above that no even harmonics appear at the output of a balance circuit when this is driven by two input signals of equal amplitude and opposite phase. The absence of even-order harmonics can also be observed in the output waveform as indicated the following theorem.

**THEOREM 2.1.** If for a signal f(t) without DC component and with a period  $T = \omega/(2\pi)$ , t following condition holds

$$f(t) = -f(t + \frac{T}{2}) \tag{2.2}$$

then no even-order harmonics of  $\omega$  appear in f(t).

This theorem can be proven easily by computing the coefficients of the Fourier series of f and taking into account equation (2.23).

We have already seen two waveforms for which equation (2.23) holds, namely the squawaveform of Figure 2.3 and the triangle waveform of Figure 2.4. The Fourier series of these two waveforms, equations (2.19) and (2.21), respectively, indeed do not contain any even harmonic

A similar theorem holds for the absence of odd-order harmonics:

**THEOREM 2.2.** If for a signal f(t) without DC component and with a period  $T = \omega/(2\pi)$ , following condition holds

$$f(t) = f(t + \frac{T}{2}) \tag{2.2}$$

then no odd-order harmonics of  $\omega$  appear in f(t).

An example of such waveform is given by the absolute value of  $\sin(\omega t)$ . The Fourier series for this waveform is given by

$$f(t) = |\sin(t)| = \frac{4}{\pi} \left( \frac{1}{2} - \frac{1}{1.3} \cos(2t) - \frac{1}{3.5} \cos(4t) - \frac{1}{5.7} \cos(6t) \right)$$
 (2.2)

and it is indeed seen that no odd-order harmonics are present. Notice that there is no component at the fundamental frequency neither (this is also an odd-order component).

## 2.4 Two-frequency excitation

In much the same way as in the previous section, the test circuit under consideration is now excited with two sinusoids  $A_1\cos(\omega_1 t)$  and  $A_2\cos(\omega_2 t)$ , both applied at the same input port. When  $A_1$  and  $A_2$  are sufficiently low, then the output spectrum contains two signals above the noise floor at the fundamental frequencies  $\omega_1$  and  $\omega_2$ , due to the circuit's linear behavior. Since in a linear circuit the superposition principle is valid, the two excitations do not produce any interfering signal. However, when  $A_1$  and  $A_2$  become larger, then, apart from the harmonics of  $\omega_1$  and  $\omega_2$ , interfering signals grow above the noise floor at the frequencies  $\omega_1 + \omega_2$ ,  $|\omega_1 - \omega_2|$ ,  $2\omega_1 + \omega_2$ ,  $|2\omega_1 - \omega_2|$ ,  $\omega_1 + 2\omega_2$  and  $|-\omega_1 + 2\omega_2|$ . The signals at  $|\omega_1 \pm \omega_2|$  are caused by second-order nonlinear behavior and are called **second-order intermodulation products**. They increase with the first power of both  $A_1$  and  $A_2$ . The other signals originate from third-order nonlinear behavior and are denoted as **third-order intermodulation products**. The signals at  $|2\omega_1 \pm \omega_2|$  increase with the square of  $A_1$  and with the first power of  $A_2$ . The signals at  $|\omega_1 \pm 2\omega_2|$  increase proportionally to  $A_1$  and with the square of  $A_2$ .

Just as in the case of the single-tone excitation, at higher input amplitudes the dependence of the intermodulation products on the input amplitudes can deviate from the dependence described in the previous paragraph. For example, when  $A_1$  is high, then the second-order intermodulation product at the frequency  $|\omega_1 \pm \omega_2|$  no longer increases proportionally with  $A_1$ . This again results in *compression* or *expansion* just as we had in the case of the single-tone excitation.

The different frequency components that can be observed at the output of the test circuit that is driven by two sinusoidal signals, are shown in Figure 2.5. It is assumed that the input signals are small enough such that the circuit behaves in a weakly nonlinear way. In order not to overload the figure, components caused by nonlinear behavior of order higher than three has not been shown. The frequencies  $|2\omega_1-\omega_2|$  and  $|\omega_1-2\omega_2|$ , where third-order intermodulation products occur, can be quite close to the fundamental frequencies, which are usually the frequencies of interest. As a result, the intermodulation products can cause severe interference. This is a problem especially in communication circuits.

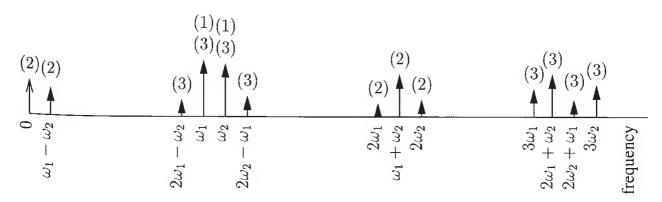


Figure 2.5: The different frequencies at the output of a weakly nonlinear analog circuit excited by two sinusoidal signals at frequencies  $\omega_1$  and  $\omega_2$ . The numbers between brackets indicate the order of nonlinearity by which the signal is determined.

In order to determine which frequency components are generated by nth-order nonlinear behavior, the following rule of thumb can be followed. Assume that k different frequencies  $\omega_1, \omega_2, \ldots, \omega_k$  are applied to the nonlinear circuit. Consider the array of frequencies

$$\omega_1, -\omega_1, \omega_2, -\omega_2, \ldots, \omega_k, -\omega_k$$

When from this array n frequencies are selected whereby a frequency may be chosen more than once, then the sum of these selected frequencies yields a frequency which can have a component caused by nth-order behavior. For example, assume that two sinusoidal input signals are applied to a circuit with frequencies  $\omega_1$  and  $\omega_2$ . In order to know which frequency components are generated by third-order nonlinear behavior, we select three frequencies from the array  $\omega_1, -\omega_1, \omega_2, -\omega_2$ . Doing so, we find that third-order nonlinear behavior gives rise to signals at the following frequencies:

$$\pm\omega_1$$
,  $\pm\omega_2$ ,  $\pm3\omega_1$ ,  $\pm3\omega_2$ ,  $\pm\omega_1\pm2\omega_2$ ,  $\pm2\omega_1\pm\omega_2$ 

The responses at negative frequencies are the complex conjugate of the responses at the corresponding positive frequencies, such that the sum of the two signals is real.

For analog circuits such as amplifiers, the intermodulation products are usually unwanted Therefore, they are denoted as intermodulation distortion. In communication circuits, these unwanted products are often denoted as spurious responses [Wein 80]. When the input amplitudes  $A_1$  and  $A_2$  are taken equal, the **second-order intermodulation distortion**  $IM_2$  and the **third-order** intermodulation distortion IM<sub>3</sub> are defined as the ratio of a second- or third-order intermodulation product to the fundamental response. When no frequency effects are taken into account and when low-distortion conditions apply, then it will be shown below that the response at  $\omega_1 + \omega_2$ is equal to the response at  $|\omega_1 - \omega_2|$ . Also, the response at  $2\omega_1 + \omega_2$  is the same as the responses at  $|2\omega_1 - \omega_2|$ ,  $2\omega_2 + \omega_1$  and  $|2\omega_2 - \omega_1|$ . Then  $IM_2$  and  $IM_3$  are given by

$$IM_2 = \left| \frac{V_{out,1,\pm 1}}{V_{out,1,0}} \right|$$
 (2.26)

$$IM_{2} = \left| \frac{V_{out,1,\pm 1}}{V_{out,1,0}} \right|$$

$$IM_{3} = \left| \frac{V_{out,2,\pm 1}}{V_{out,1,0}} \right| \stackrel{\text{low frequencies}}{=} \left| \frac{V_{out,1,\pm 2}}{V_{out,1,0}} \right|$$

$$(2.26)$$

The above definitions can be clarified by considering again the test circuit with the input output relationship given by equation (2.4). When the input signal x(t) consists of two signals of equal amplitude and with a different frequency  $\omega_1$  and  $\omega_2$ :

$$x(t) = A\cos(\omega_1 t) + A\cos(\omega_2 t) \tag{2.28}$$

then the different responses are shown in Figure 2.6. It is assumed that the amplitude A is sufficiently low such that the circuit behaves in a weakly nonlinear way. The frequencies  $\omega_1$  and  $\omega_2$  have been given the value  $2\pi \times 10MHz$  and  $2\pi \times 11MHz$ . It is seen that harmonics of  $\omega_1$  and  $\omega_2$  are present at the output as well as intermodulation products. The amplitudes of the different harmonics and intermodulation products can be computed using the formulas of Appendix A. It

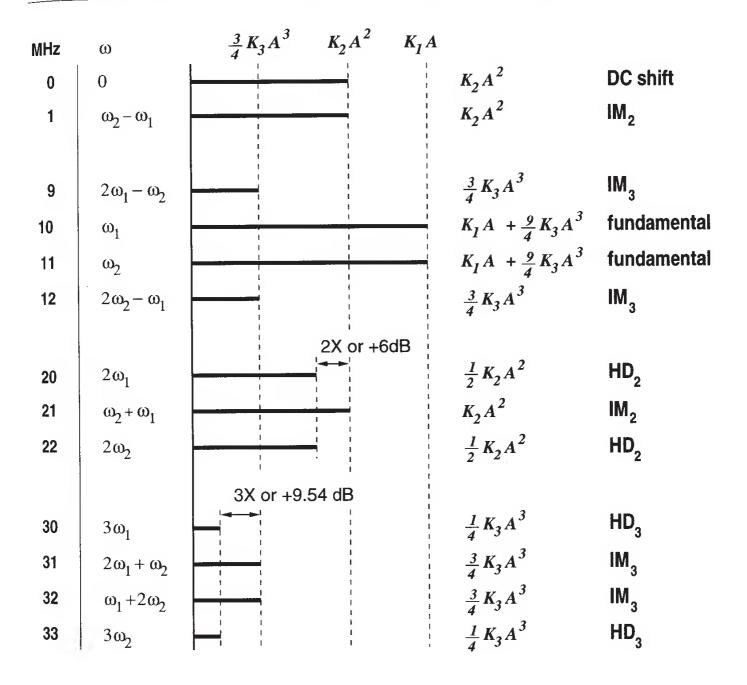


Figure 2.6: The different frequency components at the output of a weakly nonlinear circuit with the input-output relationship given by (2.4) to a combination of two sinusoidal signals with the same amplitude A. The frequency of the two input signals is 10MHz and 11MHz.

is seen that indeed the response at  $\omega_1 + \omega_2$  is equal to the response at  $|\omega_1 - \omega_2|$ . Also, the response at  $2\omega_1 + \omega_2$  is the same as the responses at  $|2\omega_1 - \omega_2|$ ,  $2\omega_2 + \omega_1$  and  $|2\omega_2 - \omega_1|$ .

From Figure 2.6 we can compute the second- and third-order intermodulation products for

our general weakly nonlinear test circuit by applying the definitions of equations (2.26) and (2.27):

$$IM_2 = \left| \frac{K_2}{K_1} \right| A \tag{2.29}$$

$$IM_3 = \frac{3}{4} \left| \frac{K_3}{K_1} \right| A^2 \tag{2.30}$$

Comparing these values with the harmonic distortion figures of equation (2.13) and (2.14), we find

$$IM_2 = 2HD_2$$
 (2.31)

$$IM_3 = 3HD_3 \tag{2.32}$$

This is an important relationship that is valid under low-distortion conditions and at low frequencies. At high input levels these relations are violated. For example, when the input amplitude is sufficiently high, then the response at the fundamental frequency also contains a third-order component (see Figure 2.6)

fundamental response = 
$$K_1V + \frac{9}{4}K_3V^3$$
 (2.33)

and the ratio of the signal at  $|2\omega_1 \pm \omega_2|$  or  $|2\omega_2 \pm \omega_1|$  with this fundamental response will not be given by equation (2.32) anymore. In addition, contributions of order higher than three will violate this relationship.

At high frequencies, when capacitors or inductors play a role, the relationships equation (2.31) and (2.32) are not satisfied neither. An explanation for this is will be given in Chapter 4 with the use of Volterra series.

The relationships (2.31) and (2.32) can be very useful when harmonics under low-distortion conditions and at low frequencies have to be measured. When harmonics are very low they can be very close to the noise floor such that they cannot be measured accurately. If two input signals with the same input amplitude are applied but with a different frequency, then a measurement of the second- and third-order intermodulation products will yield signals that are respectively 6dB and 9.54dB higher than the harmonics that were measured initially.

Compression and intercept points Just as for a single-frequency excitation, we can introduce compression points and intercept points for intermodulation when a circuit is excited with more than one frequency. Figure 2.7 depicts the wanted response of our test circuit together with one of the third-order intermodulation products. The wanted response is the fundamental response at the frequency  $\omega_1$ . The intermodulation product we are considering is the signal at  $|2\omega_1 - \omega_2|$ .

When the amplitude of the input signal at  $\omega_1$  is swept starting from a very low level, then it is seen that the output signal at  $\omega_1$  is first lost in the noise, then it increases linearly with the input amplitude, and finally it exhibits compression.

The 1dB and 3dB compression points are again defined as the input levels for which the fundamental response is 1 and 3dB lower than the extrapolation of the response at lower amplitudes which is caused by linear behavior only.

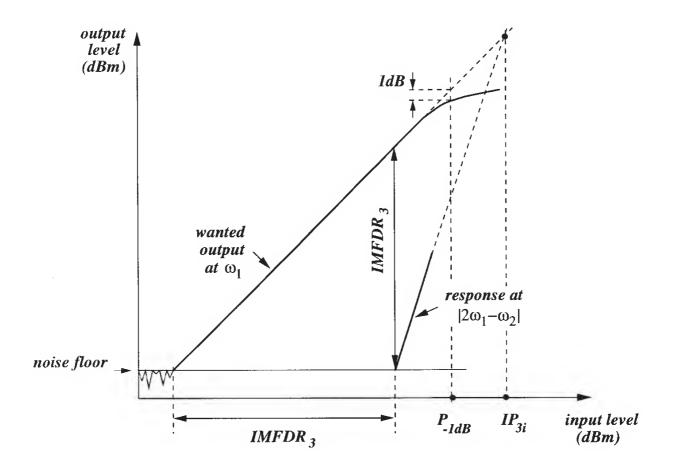


Figure 2.7: Illustration of 1-dB compression point,  $IP_{3i}$  and  $IMFDR_3$  for third-order intermodulation products.

Next, we consider the response at  $|2\omega_1 - \omega_2|$ . At low input amplitudes this response is below the noise floor. Close to the noise floor it increases with the square of the input amplitude. The third-order intercept point  $IP_{3i}$  is the input amplitude for which the asymptotes for the response at  $|2\omega_1 - \omega_2|$  and the fundamental cross. The index i of  $IP_{3i}$  is added in order to make a distinction with  $IP_{3h}$ , the third-order intercept point for harmonics.

When a second-order intermodulation product is considered, then a second-order different intercept point  $IP_{2i}$  can be defined in a similar way as  $IP_{3i}$ .

We will now compute expressions for  $IP_{2i}$  and  $IP_{3i}$  for our test circuit at low frequencies. At low frequencies the intermodulation products of the same order have the same level, such that only one  $IP_{2i}$  and  $IP_{3i}$  can be defined.

The intercept points are found by determining the input amplitude for which the second- or third-order intermodulation distortion figures are equal to one. Doing so, we find from equation (2.26)

$$IP_{2i} = \left| \frac{K_1}{K_2} \right| \tag{2.34}$$

and for  $IP_{3i}$  we use equation (2.27) to obtain

$$IP_{3i} = 2\sqrt{\left|\frac{K_1}{3K_3}\right|} \tag{2.3 5}$$

A simple relationship is found between the intercept points for harmonic and intermodulation distortion. Combining equations (2.15) and (2.34) it is found

$$IP_{2i} = \frac{IP_{2h}}{2} (2.3.6)$$

Similarly, we find with equations (2.16) and (2.35)

$$IP_{3i} = \frac{IP_{3h}}{\sqrt{3}} \tag{2.37}$$

On a logarithmic scale the ratio of  $1/\sqrt{3}$  between the two intercept points is found as a difference of  $-4.77 \, dB$ .

Intermodulation-free dynamic range Figure 2.7 shows yet another parameter: the *intermodulation-free dynamic range*  $IMFDR_3$ . This is the ratio of the largest and the smallest signal level the circuit under consideration can handle, where the intermodulation products are to remain below the noise level. The smallest signal level the circuit can handle is usually also taken equal to the noise level. The parameter  $IMFDR_3$  is read from the x-axis (at the input side) in Figure 2.7, but it can also be read from the y-axis (at the output side) as indicated. The reason is that the slope of the curve that represents the response at the fundamental frequency is 45 degrees on a double logarithmic scale.

Desensitization We consider the following experiment: the input signal  $A_1 \cos(\omega_1 t)$  is the input signal while  $A_2 \cos(\omega_2 t)$  is considered as an unwanted signal that is also applied to the input. In a communication circuit the latter signal can be a disturbing signal that is so close to the wanted signal such that any bandpass filtering before our test circuit cannot reject the disturbing signal. Assume now that the amplitude  $A_2$  is increased gradually. In a linear circuit the amplitude of the wanted response at  $\omega_1$  would be unaffected by the presence of the second signal. However, this is no longer true in a nonlinear circuit. Indeed, when  $A_2$  exceeds a critical level, then the response at  $\omega_1$  decreases. The reason is that the strong unwanted signal drives the input stage of the circuit into strongly nonlinear behavior, resulting in a lower gain for the wanted signal. This effect is called **desensitization**. Desensitization is important in communication systems where a small signal of interest (corresponding to  $A_1$ ) is influenced by a strong unwanted signal (corresponding to  $A_2$ ).

Example The need for low intermodulation distortion in communication circuits can be illustrated with the following example. Assume that the receiver front-end of Figure 1.1 has to

receive a signal that is modulated in some way with a carrier at frequency  $f_1 = 100.1 MHz$ . Suppose further that the strength of this signal is low, whereas two strong adjacent signals are present at  $f_2 = 100.2 MHz$  and at  $f_3 = 100.3 MHz$ . The signals at  $f_2$  and  $f_3$  are so close to the wanted signal that they cannot be removed by a bandpass filter around 100.1 MHz. Instead, the unwanted signals will be removed at lower frequencies further in the receiver circuitry. Due to third-order intermodulation distortion in the front-end, the two strong signals give rise to a intermodulation product at  $2f_2 - f_3 = 100.1 MHz$  which is exactly the frequency of our wanted signal. If third-order intermodulation distortion in our front-end is very high, then the intermodulation product might have a signal strength that is comparable to our wanted signal. This of course complicates the detection (demodulation) of the wanted signal.

### 2.5 Mixer definitions

A mixer performs a frequency translation either from a high-frequency input signal to a low-frequency signal or vice versa. A mixer has two inputs: a signal input and a local oscillator input. Very often, these inputs are applied at different input ports. A frequency translation to low frequencies is performed by applying to the mixer a high-frequency signal and a high-frequency local oscillator signal. The output signal of interest is at the frequency that is equal to the difference of the input signal frequency and the local oscillator frequency. This kind of frequency translation is used in receiver circuits. It is referred to as *downconversion*.

In the case of a frequency translation to high frequencies, the input signal is at low frequencies, while the frequency of the local oscillator is again high. The output signal of interest is at the frequency which is the sum or difference of the two frequencies. This frequency translation is used in transmitter circuits. It is referred to as *upconversion*.

A deep study of different mixer configurations is beyond the scope of this book. An excellent survey can be found for example in [Gilb 96]. At this point it is only important to mention that mixer configurations can be classified in two categories. In the first category, the local oscillator signal is very large such that the mixer behaves in a strongly nonlinear way with respect to the local oscillator signal. The transistors in this circuit are switched on and off at a rate determined by the local oscillator signal. Roughly speaking, this gives rise to a square-wave signal. This square wave is modulated by the input signal, and the resulting waveform contains a frequency component at the sum or difference frequency. This category is denoted as the class of *switching mixers*.

In a second class of mixers, the amplitude of the two input signals is kept fairly low, such that the mixer behaves in a weakly nonlinear way. The output signal at the sum or difference frequency is caused by second-order nonlinear behavior of the mixer. Indeed, in Figure 2.6 it is seen that second-order nonlinear behavior of a circuit that is excited by two signals of a different frequency, gives rise to a response at the sum or difference frequency. However, the situation is more complicated here, since the input signal and the local oscillator signal are usually applied at different input ports. This category of mixers is referred to as weakly nonlinear mixers.

In the literature both switching mixers [Mey 94, Sato 96, Madi 96] and weakly nonlinear mixers [Crols 95a, King 97, Borre 97] are found nowadays. In this book only weakly nonlinear

mixers will be addressed. An analytic treatment of switching mixers is only possible by making simplifications [Mey 86].

The wanted signal at the output of a mixer is the second-order intermodulation product at the sum or difference frequency, while both the linear response and intermodulation products of order higher than two are unwanted. Assume that the input signal of a mixer is a sine wave. The *conversion gain* of a mixer is the ratio of the amplitude of the wanted second-order intermodulation product at the output to the amplitude of the input signal. For a switching mixer the conversion gain is proportional to the amplitude of the local oscillator signal.

Other mixer definitions such as compression point and third-order intercept point are identical to the definitions presented in Section 2.4.

#### 2.6 Cross modulation

**Cross modulation** is the nonlinear effect whereby modulation from one carrier is transferred to another. Assume that the input x(t) to a nonlinear circuit is an amplitude-modulated signal at  $\omega_1$  together with an unmodulated carrier with frequency  $\omega_2$ :

$$x(t) = A(1 + m_1 \cos \omega_m t) \cos \omega_1 t + A \cos \omega_2 t \tag{2.38}$$

In this equation  $m_1$  is the modulation index and A is the amplitude of both signals. If the circuit has a cubic nonlinearity, then a distorted version of the modulation of the signal at  $\omega_1$  is transferred to the carrier at  $\omega_2$ .

The effect of cross-modulation can be understood by computing the response of the test circuit with the input-output relationship of equation (2.4), to the amplitude-modulated signal of equation (2.38). The response can be computed using simple algebra. Among the different terms of the response, the following two are interest:

response to AM-signal 
$$= \ldots + AK_1\cos(\omega_2 t) + 6A^3K_3\cos^2(\omega_1 t)m_1\cos(\omega_m t)\cos(\omega_2 t) + \ldots$$
 (2.39)

With the trigonometric relationships of Appendix A we find

$$\cos^2(\omega_1 t) = \frac{1}{2} + \frac{\cos(2\omega_1 t)}{2} \tag{2.40}$$

In this way, an amplitude-modulated signal is found in the response:

response to AM-signal 
$$= \ldots + AK_1 \left(1 + m_2 \cos \omega_m t\right) \cos(\omega_2 t) + \ldots$$
 (2.41)

in which the modulation index  $m_2$  is given by

$$m_2 = \frac{3K_3m_1A^2}{K_1} \tag{2.42}$$

The cross-modulation factor CM is the ratio of the two modulation indices  $m_2$  and  $m_1$ :

$$CM = \frac{m_2}{m_1} = \frac{3K_3A^2}{K_1} \tag{2.43}$$

There is a simple relationship between the cross modulation factor and third-order intermodulation distortion. Indeed, combining equations (2.30) and (2.43) yields

$$\frac{CM}{IM_3} = 4 \tag{2.44}$$

At low frequencies, the modulated signal at  $\omega_2$  is amplitude-modulated. At high frequencies, however, both amplitude and phase modulation can occur at  $\omega_2$  [Mey 72]. In order to provide quantitative measures for cross modulation at high frequencies, a knowledge of Volterra series is required. Therefore this topic will be resumed in Chapter 4.

## 2.7 Summary

In this chapter, several performance parameters have been defined that are used to quantify the nonlinear behavior of a circuit. Most definitions are valid both for weakly and strongly nonlinear behavior. When the scope of the defined parameters is limited to weakly nonlinear circuits, then at low frequencies these parameters can be expressed in terms of first-order, second-order and third-order coefficients that describe the input-output relationship of the circuit. In Chapters 4 and 5 it will be explained how these coefficients can be computed when an explicit expression for the input-output relationship cannot be computed. Also, the above definitions will be extended such that the influence of frequency can be taken into account.

## Chapter 3

# Description of nonlinearities in analog integrated circuits

#### 3.1 Introduction

Prior to the analysis of nonlinear behavior of analog circuits, it is necessary to describe the nonlinear devices that are present in analog integrated circuits. The devices most commonly used in silicon analog integrated circuits are transistors, resistors, capacitors and diodes. In the last few years self inductances and mutual inductances have been integrated as well [Ngu 90, Ash 96, Long 95, Long 95]. Their nonlinear behavior is not considered in this book.

In circuit analysis the devices mentioned above are described using an equivalent circuit. This equivalent circuit can be as simple as one circuit element (e.g. one resistor), or it consist of several circuit elements (e.g. a transistor). The elements of such equivalent circuit are nonlinear in general. The following circuit elements are used in numerical simulation of analog integrated circuits as a part of the equivalent circuit of a device:

- a nonlinear conductance: the current through this element is an algebraic function of the voltage over the element;
- a nonlinear transconductance: the current through this element is an algebraic function of a voltage other than the voltage over the element.
- a nonlinear resistance: the voltage over this element is an algebraic function of the current through this element;
- a nonlinear transresistance: the voltage over this element is an algebraic function of a current different from the current through this element;
- a multidimensional nonlinear conductance or transconductance: the current through this element is a function of more than one voltage;
- a multidimensional nonlinear resistance or transresistance: the voltage across this element is a function of more than one current;

- a nonlinear capacitance: the charge on this element is an algebraic function of the voltage across the element;
- a nonlinear transcapacitance: the charge on this element is an algebraic function of a voltage other than the voltage across the element;

These circuit elements are referred to as *basic nonlinearities*, since they are the building elements for nonlinear equivalent circuits of devices such as transistors, diodes, integrated resistors, . . . . Nonlinear voltage-controlled voltage sources and current-controlled current sources are seldom used in equivalent circuits of devices in analog integrated circuits.

In this book a nonlinear circuit is analyzed by studying excursions around a quiescent point. This means that every basic nonlinearity undergoes such excursion. A small excursion around the quiescent point of a basic nonlinearity can be described as a power series of the algebraic function that describes the basic nonlinearity. When those excursions are small enough, then this power series can be broken down after the first few terms. For very small excursions one can even do with the first term of every power series: since this term describes the linear behavior of a circuit element, this case corresponds to a linearization of the circuit.

In Chapters 4 and 5 it will become clear that a power series description of basic nonlinearities is elegant when harmonics and intermodulation products are studied that are caused by weakly nonlinear behavior.

A power series description of a basic nonlinearity contains the derivatives of the output quantity (current or voltage) with respect to the controlling quantities. These derivatives are evaluated in the quiescent point. An accurate description of a nonlinear device in terms of power series requires that the different derivatives that are considered be accurate. These derivatives are a function of the controlling quantities and of the model parameters that describe the nonlinear device. It is clear that a poor model for a nonlinear element — either due to inaccurate model equations or to an inaccurate parameter extraction — can give rise to dramatic errors on the derivatives. This is illustrated in Figure 3.1.

This figure shows a possible situation where the measured quantity is approximated by a model using a least-squares criterion. Although the error on the quantity itself might be small, it is clear that the derivative of the fitted curve seriously differs from the real derivative. This problem has been noticed by many researchers in analog design, especially with respect to the modelling of MOS transistors [Tsiv 88, Tsiv 93b, Enz 95, Groen 94, Tsiv 81].

The majority of performance parameters of an analog integrated circuit is related to the small-signal behavior of the circuit. Since the small-signal parameters are first-order derivatives, it is clear that an accurate analysis requires accurate first-order derivatives. Although many MOS models provide reasonable values for the drain current, they can give inaccurate values for some small-signal parameters. Especially the value of the output conductance is often erroneous. For distortion analysis this problem is even more severe: here, derivatives are required of order higher than one, and the deviation between a model and the real behavior tends to increase with the order of the derivative.

<sup>&</sup>lt;sup>1</sup>Usually, one does not consider nonlinear effects of order higher than three or four, hence derivatives of order higher than four do not need to be considered. In most cases one even limits to order three.

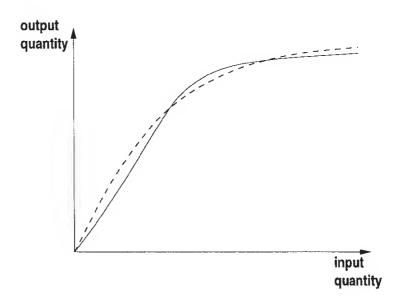


Figure 3.1: Fit of a model equation with a least-squares approximation (solid line) to measument results (dashed line).

Deviations between a model and the reality due to a poor parameter extraction can be minimized if in the parameter extraction procedure not only the deviation of the current is minimized but also the deviation on the derivatives, as suggested in [Tsiv 88]. This of course requires that the derivatives can be measured. In [Groen 94] the derivatives are computed by taking differences of the measured current. However, it is possible to directly measure the derivatives as described in [Iked 72] for the first-order derivatives and in Chapter 9 for higher-order derivatives.

Of course, efforts in the parameter extraction phase are not useful if the model itself is not adequate. Inaccuracies in the models can have several reasons. First, a model is always a simplification of the real situation. Models are kept relatively simple for several reasons: a simple model can be quickly evaluated in numerical circuit simulation, such that the computation times remain acceptable. Secondly, a simple model yields insight in the behavior of the device. Transistors, for example, are usually modeled in such a way that the current flows into one direction and that the electric fields in the transistors are one-dimensional. Due to the scaling of the devices, this assumption is more and more violated, and two-dimensional or three-dimensional effects become important. Also, equivalent circuits of devices are composed of lumped elements, whereas in reality a device is composed of distributed elements such as RC-transmission lines. An accurate simulation can only be performed by three-dimensional computer simulations. Since this is not efficient, much effort has been done in the world of device modeling to make acceptable simplifications by using empirical or semi-empirical approaches. In these approaches the twoor three-dimensional effects are broken down into simple, separate effects examined one at a time. Some simplifying assumptions are then made, which are sometimes difficult to justify, and relatively simple relationships are derived. The danger exists that oversimplification yields model equations that again describe the output quantity of the device (either a current or a voltage) relatively accurate, while the inaccuracies on the derivatives are unacceptable. This danger is especially present in the modelling of MOS transistors, as will be discussed in Chapter 7. Fortunately, only the lowest-order derivatives are required in the analyses carried out in this book and with some corrections on widely used models, insight can still be obtained in combination with a good agreement between calculations and measurements. This will also be illustrated in Chapters 9 and 8.

Devices with different operating regions are sometimes modeled with different model equations. Examples are the different model equations for the weak and strong inversion region for a MOS transistor that are used in many transistor models. The models are usually constructed in such a way that there is no "jump" in the output quantity (in this case the drain current) when switching from one model equation to another. Unfortunately, the derivatives often exhibit a discontinuity at the transition point. This can cause problems in the computation of the DC solution of a circuit or in time-domain simulations, and it will also give rise to erroneous values of the small-signal parameters in the vicinity of the transition point. Higher-order derivatives, of course suffer from the same problem. This problem is not solved in this book. It is circumvented by assuming that the bias point of a device is far enough from the transition point and that the signal swings are small enough such that the device remains in the same operating region. Of course, this problem does not occur for devices that are described with one model equation, such as the bipolar transistor. For MOS transistors, recent research efforts have resulted in the modelling of the MOS current with one equation as well [BSIM 95, Foty 96].

With the power series approach the DC solution is split from the AC part. Since it has already been pointed out that many device models do a decent job for the computation of the DC solution whereas they are inaccurate for the derivatives, one can use such model for the DC solution and a separate model for the AC solution. This has also been suggested in [Tsiv 83, Tsiv 88].

This chapter is organized as follows: in Section 3.2 the basic nonlinearities are described in terms of power series. These basic nonlinearities are the composing elements of equivalent circuits of devices such as diodes and transistors. Next, the nonlinearity of integrated resistors and capacitors is discussed in Section 3.2.2 and 3.2.3, respectively. The basic nonlinearities can be used to build transistor models, as will be mentioned in Section 3.5. Nevertheless, a detailed discussion of the different nonlinearities in bipolar and MOS transistors is quite involved. This is discussed in separate chapters, namely Chapters 6 and 7. The diode is not discussed as a separate device: its nonlinear behavior can be extracted from Chapter 6.

## 3.2 Power series description of basic nonlinearities

In the introduction an enumeration has been given of the different nonlinear circuit elements that can make part of the equivalent circuit of a device. For most applications, one can do with a more restrictive set of circuit elements: a nonlinear resistance, a nonlinear conductance, a nonlinear transconductance, a nonlinear capacitance, and a nonlinear current source depending upon several controlling voltages, often indicated as a multidimensional transconductance<sup>2</sup>. They are depicted in Figure 3.2. These nonlinearities are denoted as the *basic nonlinearities*. Their general description in terms of power series is given below. These basic nonlinearities are the

<sup>&</sup>lt;sup>2</sup>Although the term transconductance is used, one of the controlling voltages can be the voltage over the element itself.

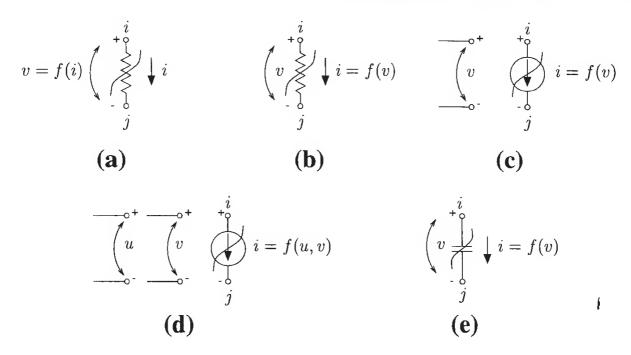


Figure 3.2: The basic nonlinearities used for the analysis of the weakly nonlinear behavior of analog integrated circuits: (a) a nonlinear resistance, (b) a nonlinear conductance, (c) a nonlinear transconductance, (d) a multidimensional transconductance and (e) a nonlinear capacitance.

building blocks of nonlinear models of devices such as transistors, as described in the subsequent sections.

Multidimensional transconductances are seldom used in distortion calculations found in literature on computations of distortion in analog integrated circuits. The reason is that most computations found in literature are simplified that much such that only one AC voltage that controls a multidimensional transconductance, differs from zero. For more complicated computations, however, one must allow that all controlling voltages of a multidimensional transconductance are different from zero. This can only be described accurately with multidimensional transconductances. It will be seen in Sections 3.2.4 and 3.2.5 that in the power series description of multidimensional transconductances terms arise that are due to the simultaneous presence of two or three controlling voltages (in AC). These terms, in fact cross-terms, do not occur in purely linear circuits, and they cannot be obtained by describing a multidimensional transconductance as the sum of one-dimensional (trans)conductances.

A description of nonlinear transcapacitances is omitted in this book. Their description is similar to the description of nonlinear capacitances. Transcapacitances are used for example in the high-frequency modelling of MOS transistors [Tsiv 88].

In the modelling of devices that occur in analog integrated circuits, the nonlinear elements without memory (no capacitor or inductor) are most often described as voltage-controlled elements. Current-controlled elements occur quite seldom. An example is a nonlinear resistance that suffers from current crowding, such as the base resistance of a bipolar transistor. In most cases, however, it is possible to construct a power series in terms of voltages for a circuit element that is originally described as a resistance. Conversion formulas are given in Section 3.2.2.

## 3.2.1 Nonlinear conductance and transconductance

For a nonlinear conductance or transconductance, the current through the element,  $i_{OUT}(t)$ , is a nonlinear function  $f_1$  of the controlling voltage  $v_{CONTR}(t)$ . For a conductance this is the voltage over the element itself, whereas for a transconductance this is a voltage elsewhere in the circuit. This function can be expanded into a power series around the quiescent point  $I_{OUT} = f_1(V_{CONTR})$ :

$$i_{OUT}(t) = f(v_{CONTR}(t)) = f(V_{CONTR} + v_{contr}(t))$$

$$= f(V_{CONTR}) + \sum_{k=1}^{\infty} \frac{1}{k!} \left. \frac{\partial^k f(v(t))}{\partial v^k} \right|_{v=V_{CONTR}} \cdot v_{contr}^k(t)$$
(3.1)

In this equation  $i_{OUT}(t)$  is the total value of the current, which is the sum of the DC and the AC current. The voltage  $v_{contr}(t)$  is the AC voltage that controls the conductance. The second term in equation (3.1) is a power series representing the AC part of the current.

When the analysis of a circuit that contains a nonlinear conductance is limited to first-, second- and third-order nonlinear behavior, then the power series in equation (3.1) can be broken down after the third term.

Defining the following coefficients

$$g_1 = \left. \frac{\partial f(v)}{\partial v} \right|_{v = V_{CONTR}} \tag{3.2}$$

$$K_{2g_1} = \frac{1}{2!} \left. \frac{\partial^2 f(v)}{\partial v^2} \right|_{v = V_{CONTR}}$$
(3.3)

$$K_{3g_1} = \frac{1}{3!} \left. \frac{\partial^3 f(v)}{\partial v^3} \right|_{v = V_{CONTR}}$$
(3.4)

in which we omitted the time dependence for simplicity, and in general

$$K_{ng_1} = \frac{1}{n!} \left. \frac{\partial^n f(v)}{\partial v^n} \right|_{v = V_{CONTR}}$$
(3.5)

leads to the expression of the AC current through the conductance

$$i_{out}(t) = g_1 \cdot v_{contr}(t) + K_{2g_1} \cdot v_{contr}^2(t) + K_{3g_1} \cdot v_{contr}^3(t) + \dots$$
 (3.6)

In this expression,  $g_1$  is the small-signal (trans)conductance of the linearized element. The coefficients in the second and third term,  $K_{2g_1}$  and  $K_{3g_1}$  are respectively the **second**- and **third-order nonlinearity coefficients** that describe the nonlinear element. Similarly, the small-signal conductance is often referred to as the first-order coefficient. The subscript for  $K_2$  and  $K_3$  is the symbol that represents the linearized element, in this case  $g_1$ . This convention will be followed for the other basic nonlinearities as well.

Sometimes it is interesting to normalize the second- and third-order nonlinearity coefficients with respect to the first-order coefficient. In that case,

$$K_{2g_1}' = K_{2g_1}/g_1 \tag{3.7}$$

is the second-order normalized nonlinearity coefficient, and

$$K_{3g_1}' = K_{3g_1}/g_1 (3.8)$$

is the *third-order normalized nonlinearity coefficient*. Normalized nonlinearity coefficients are interesting to consider since, according to equation (2.13) and (2.14) they are proportional to the harmonic distortion of the current through the nonlinear conductance as a result of a sinusoidal controlling voltage.

The units of the coefficients  $g_1$ ,  $K_{2g_1}$  and  $K_{3g_1}$  are A/V,  $A/V^2$  and  $A/V^3$ , respectively. The units for the normalized nonlinearity coefficients  $K'_{2g_1}$  and  $K'_{3g_1}$  are  $V^{-1}$  and  $V^{-2}$ , respectively.

It is important to note that the polarity of the nonlinear conductance must be taken into account. For a linear conductance it is the same which of both terminals is the positive node. If in Figure 3.2 for a linear element node i is chosen as the positive node, then the relation between the current and the controlling voltage is exactly the same as if node j were the positive node. On the other hand, if for a nonlinear conductance node i is chosen as the positive node, then the current flows from node i to j, and the current is described as a function  $i_1$  of  $v_i - v_j$ . For node j chosen as the positive node, the current flows from node j to i and is represented by a function  $i_2$  of  $v_j - v_i$ . Generally, the (nonlinear) relationship between  $i_2$  and  $v_j - v_i$  is not the same as the relation between  $i_1$  and  $v_i - v_j$ . For example, consider a transconductance that is described by the relationship

$$i_1 = \exp\left(v_i - v_i\right) \tag{3.9}$$

If the polarity is reversed, then the transconductance is described as

$$i_2 = -\exp(-(v_j - v_i))$$
 (3.10)

$$\neq \exp\left(v_j - v_i\right) \tag{3.11}$$

Developing the functions of both equations (3.10) and (3.11) into a power series, reveals that the terms of the odd powers are equal while the ones of the even powers are opposite. If, however, the function that describes the nonlinearity is an odd function, then no even powers are present in the power series expansion and the polarity can be reversed without altering the functional relationship.

#### 3.2.1.1 Example: collector current of a BJT

A simplified model of the collector current  $i_C$  of a bipolar transistor is given by

$$i_C = I_S \exp(\frac{v_{BE}}{V_t}) \tag{3.12}$$

in which  $I_S$ ,  $v_{BE}$  and  $V_t$  are the transistor saturation current, the base-emitter voltage and the thermal voltage, respectively. This is a nonlinear transconductance: the current which flows between collector and emitter is controlled by the base-emitter voltage difference.

The first derivative of the collector current with respect to the base-emitter voltage is the transistor transconductance  $g_m$ . Hence, the nonlinearity coefficients that describe the nonlinearity of the collector current are denoted by  $K_{2g_m}$  and  $K_{3g_m}$ . Using the definitions of equations (3.2) through (3.4) in conjunction with equation (3.12) yields

$$g_m = \frac{I_C}{V_t} \tag{3.13}$$

$$K_{2g_m} = \frac{I_C}{2! \, V_t^2} = \frac{g_m}{2V_t} \tag{3.14}$$

$$K_{3g_m} = \frac{I_C}{3! \, V_t^3} = \frac{g_m}{6V_t^2} \tag{3.15}$$

and in general

$$K_{ng_m} = \frac{I_C}{n! \, V_t^n} = \frac{g_m}{n! \, V_t^{n-1}} \tag{3.16}$$

The normalized nonlinearity coefficients are then given by

$$K'_{ng_m} = \frac{1}{n! \, V_t^{n-1}} \tag{3.17}$$

For a quiescent current of 1mA and with  $V_t=0.0258V$ , one obtains  $g_m=0.0387A/V$ ,  $K_{2g_m}=0.751A/V^2$  and  $K_{3g_m}=9.70A/V^3$ . The second- and third-order normalized nonlinearity coefficients are found to be  $K_{2g_m}'=19.4V^{-1}$  and  $K_{3g_m}'=250.6V^{-2}$ .

#### 3.2.1.2 Example: base current of a BJT

The base current of a bipolar transistor is an example of a one-dimensional conductance. The first-order model of the base current is given by

$$i_B = \frac{I_S}{\beta_F} \exp(\frac{v_{BE}}{V_t}) \tag{3.18}$$

in which  $\beta_F$ , the transistor beta [Lak 94] is considered as a constant. Comparing this simple model to the simple model of the collector current (equation (3.12)) it is seen that the two currents differ by a constant. Hence, the shape of the two nonlinearities is identical.

The first-order derivative of the base current with respect to the base-emitter voltage is the inverse of the small-signal resistance  $r_{\pi}$ , denoted as  $g_{\pi}$ . In this way, the nonlinearity coefficients

are

$$g_{\pi} = \frac{I_B}{V_t} \tag{3.19}$$

$$=\frac{I_C}{\beta_F V_t} \tag{3.20}$$

$$K_{2g_{\pi}} = \frac{I_B}{V_t^2} = \frac{g_{\pi}}{2V_t} \tag{3.21}$$

$$=\frac{g_m}{2\beta_E V_t} \tag{3.22}$$

$$K_{3g_{\pi}} = \frac{I_B}{6V^3} = \frac{g_{\pi}}{6V^2} \tag{3.23}$$

$$=\frac{g_m}{6\beta_E V_t^2} \tag{3.24}$$

and in general

$$K_{ng_{\pi}} = \frac{I_B}{n! V_t^n} = \frac{g_{\pi}}{n! V_t^{n-1}} \tag{3.25}$$

$$=\frac{g_m}{n!\beta_F V_t^{n-1}}\tag{3.26}$$

More advanced descriptions of the collector and base current of a bipolar transistor will be presented in Section 3.5.

#### 3.2.2 Nonlinear resistance

A nonlinear resistance is a current-controlled element: the voltage over this element is a nonlinear function f of the current through the element. This function can again be expanded into a power series around the quiescent point:  $V_{OUT} = f(I_{CONTR})$ :

$$v_{OUT}(t) = f(i_{CONTR}(t)) = f(I_{CONTR} + i_{contr}(t))$$

$$= f(I_{CONTR}) + \sum_{k=1}^{\infty} \frac{1}{k!} \left. \frac{\partial^k f(i(t))}{\partial i^k} \right|_{i=I_{CONTR}} \cdot i_{contr}^k(t)$$
(3.27)

The second term in equation (3.27) is a power series representing the AC part of the voltage over the element. Since this voltage is the output quantity, it has the subscript "out", whereas the controlling quantity is the current, which has a subscript "contr".

When the analysis of a circuit that contains a nonlinear resistance is limited to first-, secondand third-order nonlinear behavior, then the power series in equation (3.27) can be broken down after the third term. Defining the following coefficients

$$r_1 = \left. \frac{\partial f(i)}{\partial i} \right|_{i = I_{CONTR}} \tag{3.28}$$

$$K_{2r_1} = \frac{1}{2!} \cdot \frac{\partial^2 f(i)}{\partial i^2} \bigg|_{i=I_{CONTR}}$$
(3.29)

$$K_{3r_1} = \frac{1}{3!} \cdot \left. \frac{\partial^3 f(i)}{\partial i^3} \right|_{i=I_{CONTR}}$$
(3.30)

and in general

$$K_{nr_1} = \frac{1}{n!} \cdot \frac{\partial^n f(v)}{\partial i^n} \bigg|_{i=I_{CONTR}}$$
(3.31)

leads to the expression of the AC voltage

$$v_{out}(t) = r_1 \cdot i_{contr}(t) + K_{2r_1} \cdot i_{contr}^2(t) + K_{3r_1} \cdot i_{contr}^3(t) + \dots$$
 (3.32)

In this expression,  $i_{contr}(t)$  denotes the AC value of the controlling current. For a resistance, this is the current over the element, while for a transresistance, this is an AC current elsewhere in the circuit. The coefficient of the first term in the power series above,  $r_1$ , is the small-signal (trans)resistance of the linearized element. The coefficients in the second and third term,  $K_{2r_1}$  and  $K_{3r_1}$  are respectively the **second**- and **third-order nonlinearity coefficients** that describe the nonlinear resistance. Similarly, the small-signal (trans)resistance is often referred to as the **first-order nonlinearity coefficient**. The units of the coefficients  $r_1$ ,  $K_2$  and  $K_3$  are  $\Omega$ ,  $\Omega/A$  and  $\Omega/A^2$ , respectively.

Sometimes it is interesting to normalize the second- and third-order nonlinearity coefficients with respect to the first-order coefficient. In that case,  $K'_{2r_1} = K_{2r_1}/r_1$  is the second-order normalized nonlinearity coefficient, and  $K'_{3r_1} = K_{3r_1}/r_1$  the third-order normalized nonlinearity coefficient. The units for the second- and third-order normalized nonlinearity coefficients are  $A^{-1}$  and  $A^{-2}$ , respectively.

## 3.2.2.1 Conversion formulas between a voltage-controlled description and a current-controlled description

Many practical circuit elements can be described both as voltage-controlled and as current-controlled elements. Usually, the physical origin of a nonlinearity gives rise to one of the two options: for example, the value of the base resistance changes with the base current due to current crowding. This appeals for a description of the base resistance as a current-controlled element. Nevertheless, in many situations, a description in terms of a controlling voltage is mathematically correct as well, and will lead to the same results. Indeed, if, for example, a nonlinear conductance is described as

$$i = f(v) \tag{3.33}$$

then the element can be described as well as

$$v = f^{-1}(i) (3.34)$$

if at least the inverse function  $f^{-1}$  exists. In practice, one of the two descriptions (3.33) or (3.34) is an explicit relationship derived from physical considerations. The other relationship, on the other hand, cannot always be written explicitly. Nevertheless, as will be shown below, it is possible under certain conditions to find the derivatives even without having to determine that relationship explicitly. These derivatives can then be used in the power series description of the nonlinear element. The determination of the first-order derivative can be performed using the following theorem. In this theorem we denote the derivative of a function with a prime.

**THEOREM 3.1.** Let the function f be a continuous function over an interval [x, y] such that the function value for every element inside this interval is unique. If f can be differentiated in  $f^{-1}(a)$  for  $a \in [x, y]$  and  $f'(f^{-1}(a)) \neq 0$ , then  $f^{-1}$  can be differentiated in a and

$$(f^{-1})'(a) = \frac{1}{f'(f^{-1}(a))}$$
 (3.35)

This theorem establishes a relationship between the derivative of the inverse function and the derivative of the function itself. This theorem will be useful for the derivation of the power series coefficients of the inverse function.

If the conditions stated in Theorem 3.1 are not only fulfilled for f but also for f' and f'', which is the second derivative, then one can easily find

$$(f^{-1})''(a) = -\frac{f''(f^{-1}(a))}{(f'(f^{-1}(a)))^3}$$
(3.36)

$$(f^{-1})'''(a) = \frac{1}{(f'(f^{-1}(a)))^4} \cdot \left(-f'''(f^{-1}(a)) + 3\frac{(f''(f^{-1}(a)))^2}{f'(f^{-1}(a))}\right)$$
(3.37)

These expressions can now be used to compute the nonlinearity coefficients of a voltage-controlled description from the nonlinearity coefficients of a current-controlled description. Assume that a current-controlled description is given by

$$v = r_{ac} \cdot i + K_{2r_{ac}} \cdot i^2 + K_{3r_{ac}} \cdot i^3 + \dots$$
 (3.38)

while the equivalent voltage-controlled description is given by

$$i = g_{ac} \cdot v + K_{2g_{ac}} \cdot i^2 + K_{3g_{ac}} \cdot i^3 + \dots$$
 (3.39)

then, using Theorem 3.1 and equations (3.36) and (3.37), one obtains

$$g_{ac} = \frac{1}{r_{ac}} \tag{3.40}$$

$$K_{2g_{ac}} = -\frac{K_{2r_{ac}}}{r_{ac}^3} \tag{3.41}$$

$$K_{3g_{ac}} = \frac{1}{r_{ac}^4} \left( -K_{3r_{ac}} + 2 \frac{\left(K_{2r_{ac}}\right)^2}{r_{ac}} \right) \tag{3.42}$$

and, conversely,

$$r_{ac} = \frac{1}{g_{ac}} \tag{3.43}$$

$$K_{2r_{ac}} = -\frac{K_{2g_{ac}}}{g_{ac}^3} \tag{3.44}$$

$$K_{3r_{ac}} = \frac{1}{g_{ac}^4} \left( -K_{3g_{ac}} + 2 \frac{\left(K_{2g_{ac}}\right)^2}{g_{ac}} \right) \tag{3.45}$$

These conversion formulas will be used for example in Chapter 6 for computations which involve the nonlinearity of the base resistance.

#### 3.2.2.2 Difference between DC and AC resistance

For a nonlinear resistance it is interesting to note that the DC value is different from the AC value. This is illustrated for a nonlinear resistor in Figure 3.3, which shows the voltage over a nonlinear resistor as a function of the current through the resistor. The DC and the AC resistance

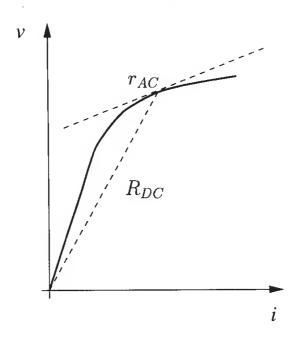


Figure 3.3: Voltage as a function of current for a nonlinear resistor. The DC and AC resistances are given by the indicated slopes.

are indicated on the figure as well. The DC resistance is the ratio of the voltage over the resistor and the current that flows through the element. When this voltage is evaluated as a function of current, then the slope — or derivative — of this curve is the AC resistance. The secondand third-order derivatives of this curve are equal to  $2!K_{2r_{AC}}$  and  $3!K_{3r_{AC}}$ , respectively. It is important to note that for a linear resistor there is no difference between the AC and DC value.

Indeed, for a linear resistor, the curve that represents the voltage as a function of current is a straight line. Hence, the ratio of the voltage and the current is the same as the slope of this curve.

In a similar way as for resistances, a difference can be made between the DC value and the AC value of a nonlinear conductance.

#### 3.2.3 Nonlinear capacitance

Capacitors in analog integrated circuits can be parasitic or they can be wanted components. In the first case, they occur for example in the equivalent circuit of a transistor. Integrated capacitors on the other hand, are wanted components. Both kinds exhibit nonlinear behavior.

For a nonlinear capacitor the charge is considered as the controlled quantity, and the voltage over the capacitor is the controlling quantity. The nonlinear relationship between the charge  $q_{OUT}$  and the voltage  $v_{CONTR}$  over the capacitor is given by

$$q_{OUT}(t) = f(v_{CONTR}(t)) = f(V_{CONTR} + v_{contr}(t))$$

$$= f(V_{CONTR}) + \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k f(v(t))}{\partial v^k} \Big|_{v=V_{CONTR}} \cdot v_{contr}^k(t)$$
(3.46)

The power series in equation (3.46) represents the AC part of the charge upon the capacitor. Defining the *first*, *second*- and *third-order nonlinearity coefficients* as

$$C_1 = \left. \frac{\partial f(v)}{\partial v} \right|_{v = V_{CONTR}} \tag{3.47}$$

$$K_{2_{C_1}} = \frac{1}{2!} \cdot \frac{\partial^2 f(v)}{\partial v^2} \Big|_{v=V_{CONTR}}$$
 (3.48)

$$K_{3C_1} = \frac{1}{3!} \cdot \frac{\partial^3 f(v)}{\partial v^3} \bigg|_{v=V_{GONTR}}$$
(3.49)

the AC part of the capacitor charge is written as

$$q_{out} = C_1 \cdot v_{contr}(t) + K_{2C_1} \cdot v_{contr}^2(t) + K_{3C_1} \cdot v_{contr}^3(t) + \dots$$
 (3.50)

The first-order coefficient  $C_1$  represents the small-signal value of the linearized element. The units for  $C_1$ ,  $K_{2C}$  and  $K_{3C}$  are respectively F, F/V and  $F/V^2$ .

units for  $C_1$ ,  $K_{2_{C_1}}$  and  $K_{3_{C_1}}$  are respectively F, F/V and  $F/V^2$ . Sometimes it is interesting to normalize the second- and third-order nonlinearity coefficients with respect to the first-order coefficient. In that case,  $K'_{2g_1} = K_{2_{C_1}}/C_1$  is the second-order normalized nonlinearity coefficient, and  $K'_{3_{C_1}} = K_{3_{C_1}}/C_1$  the third-order normalized coefficient of the nonlinear capacitor. The units for the second- and third-order normalized nonlinearity coefficients are  $V^{-1}$  and  $V^{-2}$ , respectively.

The AC current  $i_{out}(t)$  through a capacitor is found by taking the time-derivative of the AC charge upon the capacitor:

$$i_{out}(t) = \frac{d}{dt} \left( C_1 \cdot v_{contr}(t) + K_{2C_1} \cdot v_{contr}^2(t) + K_{3C_1} \cdot v_{contr}^3(t) + \dots \right)$$
(3.51)

#### 3.2.3.1 Example: diffusion capacitance

As an example, consider the simplified model of the diffusion capacitance between the base and emitter of a bipolar transistor. The nonlinear relationship between the capacitor charge  $Q_D$  and the voltage  $v_{BE}$  over the element is given by [Getr 76, Anto 88, Lak 94]:

$$Q_D = \tau_F i_C = \tau_F I_S \exp(\frac{v_{BE}}{V_t})$$
(3.52)

in which  $\tau_F$  is the transit time,  $i_C$  is the collector current,  $I_S$  is the saturation current of the transistor and  $V_t$  is the thermal voltage. Noting the first derivative of the charge with respect to the controlling voltage as  $C_D$ , one obtains the following nonlinearity coefficients:

$$C_D = \frac{\tau_F I_C}{V_t} = \tau_F g_m \tag{3.53}$$

$$K_{n_{C_D}} = \frac{\tau_F I_C}{n! V_t^n} = \frac{\tau_F g_m}{n! V_t^{n-1}}$$
(3.54)

For a collector current of 1mA and a forward transit time of 50ps, the following values are obtained at T=300K:  $C_D=1.94pF$ ,  $K_{2C_D}=37.6pF/V$  and  $K_{3C_D}=485pF/V^2$ .

#### 3.2.4 Two-dimensional transconductance

A two-dimensional transconductance is an element the current of which is controlled by two different voltages. In other words, the current  $i_{OUT}(t)$  is a function f of two voltages  $u_{CONTR}$  and  $v_{CONTR}$ , which can be expressed in terms of AC values using a two-dimensional power series expansion around the quiescent point  $I_{OUT} = f(U_{CONTR}, V_{CONTR})$ :

$$i_{OUT}(t) = f(u_{CONTR}(t), v_{CONTR}(t)) = f(U_{CONTR} + u_{contr}(t), V_{CONTR} + v_{contr}(t))$$
$$= f(U_{CONTR}, V_{CONTR}) +$$

$$\sum_{m=1}^{\infty} \sum_{n=0}^{m} \left[ \frac{\partial^{m} f(u,v)}{\partial u^{n} \partial v^{m-n}} \middle| \begin{array}{c} u = U_{CONTR} \\ v = V_{CONTR} \end{array} \right. \cdot \frac{u_{contr}^{n}(t)}{n!} \cdot \frac{v_{contr}^{m-n}(t)}{(m-n)!} \right]$$
(3.55)

The AC part of the current corresponds to the second term of equation (3.55), which is a twodimensional power series. This series can be split into three series  $i_1$ ,  $i_2$  and  $i_3$ , each corresponding to a part of the total AC current. The first two series,  $i_1$  and  $i_2$  contain powers of one single voltage and so they are similar to the series described in Section 3.2.1:

$$i_1 = g_1 \cdot u_{contr} + K_{2g_1} \cdot u_{contr}^2 + K_{3g_1} \cdot u_{contr}^3 + \dots$$
 (3.56)

and

$$i_2 = g_2 \cdot v_{contr} + K_{2g_2} \cdot v_{contr}^2 + K_{3g_2} \cdot v_{contr}^3 + \dots$$
 (3.57)

The time dependence of  $u_{contr}$  and  $v_{contr}$  has been omitted for simplicity. The third series,  $i_3$ , contains nothing but cross-terms, which are terms that contain a nonzero power of both  $u_{contr}$  and  $v_{contr}$ :

$$i_{3} = K_{2g_{1}\&g_{2}} \cdot u_{contr} \cdot v_{contr} + K_{3g_{1}\&g_{2}} \cdot u_{contr}^{2} \cdot v_{contr} + K_{3g_{1}\&2g_{2}} \cdot u_{contr} \cdot v_{contr}^{2} + \dots$$
(3.58)

The meaning of the subscripts in the *nonlinearity coefficients* defined above is as follows. Suppose that the first-order derivative of the total current with respect to u and v are respectively  $g_1$  and  $g_2$ . Then a coefficient like  $K_{m_{jg_1}\&(m-j)g_2}$  with m and j positive integers and m>j means

$$K_{m_{jg_1}\&(m-j)g_2} = \frac{\partial^m f(u,v)}{\partial u^j \partial v^{m-j}} \cdot \frac{1}{j!} \cdot \frac{1}{(m-j)!}$$
(3.59)

If j or (m-j) are equal to one, then they are usually omitted as a subscript, like in  $K_{2g_1\&g_2}$ . Note that the derivatives are evaluated for  $u=U_{CONTR}$  and  $v=V_{CONTR}$ . Just as with the above basic nonlinearities, one can again consider normalized nonlinearity coefficients by taking the appropriate ratios.

#### 3.2.4.1 Example: two-dimensional collector current

In order to clarify the above definitions, consider the simple equation for the collector current of a bipolar transistor including the Early effect:

$$i_C = I_S \exp(\frac{v_{BE}}{V_t}) \left( 1 + \frac{v_{CE}}{V_{AF}} \right) \tag{3.60}$$

in which  $V_{AF}$  denotes the Early voltage for the forward active mode of operation and  $v_{CE}$  is the collector-emitter voltage. Clearly, the collector current is a function of two controlling voltages. The first derivatives with respect to the controlling voltages  $v_{BE}$  and  $v_{CE}$  are the transconductance  $g_m$  and the output conductance  $g_o$ , respectively. Using these symbols, the AC value of the collector current can be written as the sum of a series containing powers of  $v_{be}$  only, a series depending only on  $v_{ce}$  and a series with nothing but cross-terms:

$$i_{c} = g_{m} \cdot v_{be} + K_{2g_{m}} \cdot v_{be}^{2} + K_{3g_{m}} \cdot v_{be}^{3} + \dots + g_{o} \cdot v_{ce} + K_{2g_{o}} \cdot v_{ce}^{2} + K_{3g_{o}} \cdot v_{ce}^{3} + \dots + K_{2g_{m} \& g_{o}} \cdot v_{be} \cdot v_{ce} + K_{3g_{m} \& 2g_{o}} \cdot v_{be} \cdot v_{ce}^{2} + \dots$$

$$(3.61)$$

in which the time dependence has been omitted for clarity. The meaning of the coefficients in this power series is as discussed above and in Section 3.2.1. The definition of the coefficients is repeated for clarity in Table 3.1.

The nonlinearity coefficients  $K_{2g_m}$  and  $K_{3g_m}$  have already been discussed earlier in the example of the one-dimensional collector current (see Section 3.2.1). The coefficient  $g_o$  is the output conductance and is given by  $I_C/V_{AF}$ . The coefficients  $K_{2g_o}$  and  $K_{3g_o}$  are zero: since the

$g_m$	$\frac{\partial i_C}{\partial v_{BE}}$	$K_{3g_o}$	$\frac{1}{6} \frac{\partial^3 i_C}{\partial v_{CE}^3}$
$K_{2g_m}$	$\left  \frac{1}{2} \frac{\partial^2 i_C}{\partial v_{BE}^2} \right $	$K_{2_{g_m}\&g_o}$	$\frac{\partial^2 i_C}{\partial v_{BE} \partial v_{CE}}$
$K_{3g_m}$	$\frac{1}{6} \frac{\partial^3 i_C}{\partial v_{BE}^3}$	$K_{3_{2g_m\&g_o}}$	$\frac{1}{2} \frac{\partial^3 i_C}{\partial v_{BE}^2 \partial v_{CE}}$
$g_o$	$rac{\partial i_C}{\partial v_{CE}}$	$K_{3_{g_m}\&2g_o}$	$\frac{1}{2} \frac{\partial^3 i_C}{\partial v_{BE} \partial v_{CE}^2}$
$K_{2g_o}$	$\frac{1}{2} \frac{\partial^2 i_C}{\partial v_{CE}^2}$		

Table 3.1: Definition of the nonlinearity coefficients of the collector current of a bipolar transistor.

simple model for the collector current assumes a linear dependence on the collector-emitter voltage difference, the higher-order derivatives of  $i_C$  with respect to  $v_{CE}$  are zero. The coefficients of the cross-terms in equation (3.61) are given by

$$K_{2g_m \& g_o} = \frac{g_m}{V_{AE}} \tag{3.62}$$

$$K_{2g_{m}\&g_{o}} = \frac{g_{m}}{V_{AF}}$$

$$K_{32g_{m}\&g_{o}} = \frac{g_{m}}{2V_{t}V_{AF}}$$
(3.62)

$$K_{3g_m\&2g_o} = 0 (3.64)$$

For an Early voltage  $V_{AF}$  of 50V and a collector current of 1mA, one finds  $g_o=2\times10^{-5}A/V$ ,  $K_{2_{g_m\&g_o}} = 7.75 \times 10^{-4} \, A/V^2 \text{ and } K_{3_{2g_m\&g_o}} = 0.015 \, A/V^3.$ 

A more advanced model of the collector current than the one from equation (3.60) will be discussed in Chapter 6.

#### 3.2.5 Three-dimensional transconductance

A three-dimensional transconductance is a current source that is controlled by three voltages. In Other words, the current is a function f of three voltages u(t), v(t) and w(t). Using a power series expansion around the quiescent value, the total value of the current can be split into a quiescent part  $I_{OUT} = f(U_{CONTR}, V_{CONTR}, W_{CONTR})$  and an AC part. This AC part is given by

$$i_{OUT}(t) = f(u_{CONTR}(t), v_{CONTR}(t), w_{CONTR}(t))$$

$$= f(U_{CONTR} + u_{contr}(t), V_{CONTR} + v_{contr}(t), W_{CONTR} + w_{contr}(t))$$

$$= f(U_{CONTR}, V_{CONTR}, W_{CONTR}) +$$

$$\sum_{k=1}^{\infty} \sum_{i=0}^{k} \sum_{j=0}^{k-i} \left[ \frac{\partial^{k} f(u, v, w)}{\partial u^{i} \partial v^{j} \partial w^{k-i-j}} \cdot \frac{u_{contr}^{i}(t)}{i!} \cdot \frac{v_{contr}^{j}(t)}{j!} \cdot \frac{w_{contr}^{k-i-j}(t)}{(k-i-j)!} \right]$$
(3.65)

In this series the derivatives are evaluated for  $u = U_{CONTR}$ ,  $v = V_{CONTR}$  and  $w = W_{CONTR}$ . The AC current can be split into distinct parts: first, there are three power series that only contain powers of one single voltage. These series correspond to a one-dimensional nonlinear conductance as discussed in Section 3.2.1. Next, there are three power series containing only cross-terms in exactly two voltages, similar to the two-dimensional series described in Section 3.2.4. Finally there is a power series that only contains cross-terms in three voltages at the same time. When only nonlinear effects up to the third order are considered, then from this last series only the first term of this series is taken into account. This series implies the introduction of the following nonlinearity coefficients:

$$K_{m_{jg_1}\&kg_2\&(m-j-k)g_3} = \frac{\partial^m f(u,v,w)}{\partial u^j \partial v^k \partial w^{m-j-k}} \cdot \frac{1}{j!} \cdot \frac{1}{k!} \frac{1}{(m-j-k)!}$$
(3.66)

When the positive integers j, k or (m-j-k) are equal to one, then they are omitted in the notation, like in  $K_{3g_1\&g_2\&g_3}$ .

#### 3.2.5.1 Example: drain current of a MOS transistor

The meaning of the newly defined coefficients is illustrated with a simple model for the drain current of an n-MOS transistor in saturation. Taking into account bulk effect and Early effect, the drain current is given by

$$i_D = \frac{\beta}{2} (v_{GS} - V_T)^2 (1 + \lambda v_{DS})$$
 (3.67)

with

$$V_T = V_{TO} + \gamma \left( \sqrt{v_{SB} + \phi} - \sqrt{\phi} \right) \tag{3.68}$$

and

$$\beta = K_P \frac{W}{L} \tag{3.69}$$

The parameters  $\lambda$ ,  $\gamma$  and  $\phi$  are the channel-length modulation factor, the body-effect coefficient and the surface inversion potential, respectively. These parameters will be discussed further in Chapter 7.

The first derivatives of the current with respect to the controlling voltages  $v_{GS}$ ,  $v_{SB}$  and  $v_{DS}$  are the small-signal parameters  $g_m$ ,  $g_{mb}$  and  $g_o$ . The symbols of these parameters will serve as the subscript of the different nonlinearity coefficients of order two and three. Using the notations introduced in equations (3.5), (3.59) and (3.66), the AC value of the drain current is given by

$$i_{d} = g_{m} \cdot v_{gs} + K_{2g_{m}} \cdot v_{gs}^{2} + K_{3g_{m}} \cdot v_{gs}^{3} + \dots + g_{o} \cdot v_{ds} + K_{2g_{o}} \cdot v_{ds}^{2} + K_{3g_{o}} \cdot v_{ds}^{3} + \dots - g_{mb} \cdot v_{sb} - K_{2g_{mb}} \cdot v_{sb}^{2} - K_{3g_{mb}} \cdot v_{sb}^{3} + \dots + K_{2g_{m} \& g_{mb}} \cdot v_{gs} \cdot v_{sb} + K_{32g_{m} \& g_{mb}} \cdot v_{gs}^{2} \cdot v_{sb} + K_{3g_{m} \& 2g_{mb}} \cdot v_{gs} \cdot v_{sb}^{2} + \dots + K_{2g_{m} \& g_{o}} \cdot v_{gs} \cdot v_{ds} + K_{32g_{m} \& g_{o}} \cdot v_{gs}^{2} \cdot v_{ds} + K_{3g_{m} \& 2g_{o}} \cdot v_{gs} \cdot v_{ds}^{2} + \dots + K_{2g_{mb} \& g_{o}} \cdot v_{sb} \cdot v_{ds} + K_{32g_{mb} \& g_{o}} \cdot v_{sb}^{2} \cdot v_{ds} + K_{3g_{mb} \& 2g_{o}} \cdot v_{sb} \cdot v_{ds}^{2} + \dots + K_{3g_{m} \& g_{mb} \& g_{o}} \cdot v_{gs} \cdot v_{sb} \cdot v_{ds} + \dots$$

$$(3.70)$$

In this equation, the first three lines correspond to series that describe the dependence of the AC drain current on one single AC voltage. The following three lines represent the variation of the current when two voltages change at the same time. Hereby we make use of the nonlinearity coefficients that have been defined to describe two-dimensional conductances. The last line in equation (3.70) describes the variation of the current with the three controlling voltages at the same time.

It is seen that the terms that depend on  $v_{sb}$  only, are preceded by a minus sign. The reason is that the bulk transconductance  $g_{mb}$  is usually represented as a controlled source flowing from the source to the drain, which is opposite to the direction of the source  $g_m \, v_{gs}$ . This yields a positive value for  $g_{mb}$ . Hence,  $g_{mb}$  is the first-order derivative times -1. For consistency, the coefficients  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  are adjusted in the same way.

Equation (3.70) reveals that a description of the second- and third-order nonlinearity of the three-dimensional drain current in general requires sixteen second- and third-order coefficients. However, for the simple drain current model of equation (3.67) the coefficients that are computed by deriving twice with respect to  $v_{DS}$  are zero because the current is linearly dependent on  $v_{DS}$ . Also, coefficients obtained by deriving three times with respect to  $v_{GS}$  are zero because of the quadratic dependency of the drain current on  $v_{GS}$ .

Table 3.2 lists the expressions for the coefficients that describe the nonlinearity of the drain current according to equation (3.70). The expressions have been derived using the simple square-law model of equation (3.67). The table also includes numerical values for an n-MOS transistor. The parameters of this transistor are:  $W=10\mu m,~L=3\mu m,~V_{GS}=1.45V,~V_{DS}=2V,~V_{SB}=1.5V,~V_{TO}=0.9V,~\gamma=0.3V^{1/2},~\phi=0.7V,~\lambda=0.019V^{-1},~KP=5\times10^{-5}A/V^2.$  With these parameters, an effective threshold voltage of 1.09V and a drain current of  $10.96\mu A$  are found.

Table 3.2: The first-, second- and third-order coefficients for the description of the nonlinear drain (AC) current, using the model of equation (3.67).

1		
$g_m$	$\beta (1 + \lambda V_{DS})(V_{GS} - V_T)$	$6.16 \times 10^{-5}  A/V$
$K_{2g_m}$	$\frac{\beta}{2}(1+\lambda V_{DS})$	$8.65 \times 10^{-5} A/V^2$
$K_{3g_m}$	0	$0.0   A/V^3$
$g_{mb}$	$\frac{\beta \left(1 + \lambda \ V_{DS}\right) \left(V_{GS} - V_{T}\right) \ \gamma}{2 \left(\phi + V_{SB}\right)^{1/2}}$	$6.23 \times 10^{-6}  A/V$
$K_{2g_{mb}}$	$-\frac{\beta \gamma \left(1+\lambda V_{DS}\right) \left(V_{GS}-V_{TO}+\gamma \sqrt{\phi}\right)}{8 \left(\phi+V_{SB}\right)^{3/2}}$	$-1.59 \times 10^{-6}  A/V^2$
$K_{3g_{mb}}$	$\frac{\beta \gamma \left(1 + \lambda V_{DS}\right) \left(V_{GS} - V_{TO} + \gamma \sqrt{\phi}\right)}{16 \left(\phi + V_{SB}\right)^{5/2}}$	$3.16 \times 10^{-7}  A/V^3$
$g_o$	$\frac{\beta}{2}\lambda \left(V_{GS}-V_{T}\right)^{2}$	$2.01 \times 10^{-7}  A/V$
$K_{2g_o}$	0	$0.0   A/V^2$
$K_{3g_o}$	0	$0.0$ $A/V^3$
$K_{2_{g_m} \& g_{mb}}$	$-\frac{\beta \left(1+\lambda \ V_{DS}\right) \gamma}{2\sqrt{\phi+V_{SB}}}$	$-1.74 \times 10^{-5} A/V^2$
$K_{2_{g_m\&g_o}}$	$\beta \lambda \ (V_{GS} - V_T)$	$1.13 \times 10^{-6}  A/V^2$
$K_{2_{g_{mb}}\&g_o}$	$-\frac{\beta \lambda \gamma (V_{GS} - V_T)}{2\sqrt{\phi + V_{SB}}}$	$-1.14 \times 10^{-7} A/V^2$
$K_{3_{2g_m\&g_{mb}}}$	0	$0.0   A/V^3$
$K_{3_{g_m\&2g_{mb}}}$	$\frac{\beta \gamma \left(1 + \lambda V_{DS}\right)}{8 \left(\phi + V_{SB}\right)^{3/2}}$	$1.99 \times 10^{-6}  A/V^3$
$K_{3_{2g_m}\&g_o}$	$\frac{\beta\lambda}{2}$	$1.58 \times 10^{-6}  A/V^3$
$K_{3_{g_m\&2g_o}}$	0	$0.0$ $A/V^3$

$K_{3_{2g_{mb}}\&g_o}$	$\frac{\beta \lambda \gamma \left(V_{GS} - V_{TO} + \gamma \sqrt{\phi}\right)}{8 \left(\phi + V_{SB}\right)^{3/2}}$	$2.91 \times 10^{-8}  A/V^3$
$K_{3_{g_{mb}}\&2g_{o}}$	0	$0.0   A/V^3$
$K_{3_{g_m}\&g_{mb}\&g_o}$	$-\frac{\beta \lambda \gamma}{2\sqrt{\phi + V_{SB}}}$	$-3.20 \times 10^{-7}  A/V^3$

For transistor models that are more advanced, the analytic expressions for the different derivatives of the drain current are very complicated. The expressions listed above can seriously differ from the more exact values. This will be discussed in Chapter 7.

#### **Tracking nonlinearities** 3.2.6

In Section 3.2.1 we saw that the first-order expressions of the collector current and the base current of a bipolar transistor as given in equation (3.12) and (3.18), only differ by a constant factor. In general, if a nonlinear conductance or transconductance is described by the equation

$$i_{OUT1} = g(v_{CONTR1}) (3.71)$$

and a second conductance is described by

$$i_{OUT2} = ag(v_{CONTR2}) \tag{3.72}$$

in which a is a constant, then it is said that the two nonlinearities "track". The reason for this nomenclature can be explained as follows. If for both nonlinearities the current through the element is plotted versus the controlling voltage on a logarithmic scale, then the two curves are identical, apart from a translation, as shown in Figure 3.4. For the determination of the nonlinearity coefficients, the relationship (3.71) is developed into a power series around its quiescent point, after which the AC part can be identified with

$$i_{out1} = g_1 v_{contr1} + K_{2g_1} v_{contr1}^2 + K_{3g_1} v_{contr1}^3 + \dots$$
 (3.73)

in which, according to the above definitions,

$$g_1 = \frac{di_{OUT1}}{dv_{CONTR1}} \tag{3.74}$$

$$K_{2g_1} = \frac{1}{2} \frac{d^2 i_{OUT1}}{dv_{CONTR1}^2} \tag{3.75}$$

$$g_{1} = \frac{di_{OUT1}}{dv_{CONTR1}}$$

$$K_{2g_{1}} = \frac{1}{2} \frac{d^{2}i_{OUT1}}{dv_{CONTR1}^{2}}$$

$$K_{3g_{1}} = \frac{1}{6} \frac{d^{3}i_{OUT1}}{dv_{CONTR1}^{3}}$$
(3.75)

Doing the same for the relationship equation (3.72), we obtain

$$i_{out2} = g_2 \cdot v_{contr2} + K_{2g_2} \cdot v_{contr2}^2 + K_{3g_2} \cdot v_{contr2}^3 + \dots$$
 (3.77)

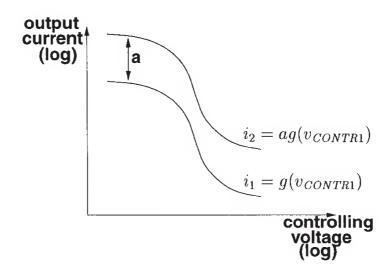


Figure `.4: The input-output relationship of two nonlinearities that track.

with

$$g_2 = \frac{di_{OUT2}}{dv_{CONTR2}} \tag{3.78}$$

$$g_{2} = \frac{di_{OUT2}}{dv_{CONTR2}}$$

$$K_{2g_{2}} = \frac{1}{2} \cdot \frac{d^{2}i_{OUT2}}{dv_{CONTR2}^{2}}$$

$$K_{3g_{2}} = \frac{1}{6} \cdot \frac{d^{3}i_{OUT2}}{dv_{CONTR2}^{3}}$$
(3.78)
$$(3.79)$$

$$K_{3g_2} = \frac{1}{6} \cdot \frac{d^3 i_{OUT2}}{dv_{CONTR2}^3} \tag{3.80}$$

If the quiescent value of the controlling voltage is identical for both nonlinearities, then it is clea from Figure 3.4, that the derivatives that determine the above nonlinearity coefficients only diffe by a factor a, which yields

$$g_2 = a \, g_1 \tag{3.81}$$

$$K_{2q_2} = a K_{2q_1} (3.82)$$

$$K_{3g_2} = a \, K_{3g_1} \tag{3.82}$$

The normalized nonlinearity coefficients can now be derived. For the nonlinear conductance described by equation (3.71) we find

$$K_{2g_1}' = \frac{\frac{d^2 i_{OUT1}}{dv_{CONTR1}^2}}{2g_1} \tag{3.84}$$

and

$$K_{3g_1}' = \frac{\frac{d^3 i_{OUT_1}}{dv_{CONTR1}^3}}{6g_1} \tag{3.85}$$

If for the second nonlinearity, described by equation (3.72), the quiescent value of the controlling voltage is the same as for the first nonlinearity, then we find for the normalized nonlinearity coefficients

$$K_{2q_2}' = K_{2q_1}' (3.86)$$

$$K_{3q_2}' = K_{3q_1}' (3.87)$$

Here we come to the important conclusion that tracking nonlinearities with identical quiescent conditions have the same normalized nonlinearity coefficients, or, in other words, they produce the same distortion.

The first-order expressions of the base and the collector current of a bipolar transistor were the first example we saw of tracking nonlinearities. Tracking nonlinearities will occur later in this book as well. The concept of tracking nonlinearities can, of course, be generalized to other basic nonlinearities as well.

#### 3.3 Integrated resistors

In bipolar processes and analog CMOS processes integrated resistors of acceptable quality can be realized in different ways. For each of the possible implementations, foundries specify a sheet resistance, the absolute accuracy, a temperature coefficient, a breakdown voltage and a voltage coefficient. The voltage coefficient indicates how much the value of the resistance varies when the voltage over the resistor changes. This variation is normalized to the nominal value of the resistor. In this way, the voltage coefficient is expressed in percent per Volt.

With the specification of a voltage coefficient no distinction is made between the DC and AC value of a resistance. In Section 3.2.2.2 it has been pointed out that there is a difference between the AC and DC value when the resistor is nonlinear. The voltage coefficients that are computed in this book address the change of the AC resistance and not the DC resistance, whereas voltage coefficients of integrated resistors as they are specified by a foundry, usually address the DC value. Moreover, the specification of the nonlinearity of a resistor by means of a voltage coefficient is incomplete. Only one value of the voltage coefficient is specified at a fixed DC voltage over the resistor.

Voltage coefficients for resistors range from 0.5%/V for diffused resistors fabricated with the emitter diffusion (bipolar process) or with the source and drain diffusion (CMOS process), to 0.02%/V for polysilicon resistors [Lak 94].

Resistors that are isolated from their substrate by means of a depletion layer have a significant voltage coefficient. In the next section, this voltage coefficient will be estimated in terms of the geometry of the resistor and of the process parameters. Also, the voltage coefficient at bias voltages different from zero will be evaluated and the third-order nonlinearity coefficient will be computed.

A voltage coefficient could as well be defined for AC values of an element. For a voltage-controlled element, such as a nonlinear conductance, this voltage coefficient indicates the relative variation of the AC conductance with the applied voltage. Since the AC conductance is nothing

H

else but the first derivative of the voltage-current relationship that describes the conductant the variation of the AC conductance corresponds to the second derivative of the voltage-current relationship. If this variation is normalized with respect to the AC conductance, then it is cluthat the voltage coefficient of an AC conductance  $g_{ac}$  is equal to two times the second-order normalized nonlinearity coefficient  $K'_{2g_{ac}}/2$ . The factor two arises from the fact that  $K_{2g_{ac}}$  only half of the second-order derivative.

#### 3.3.1 Nonlinearity coefficients of an implanted or diffused resistor

In this section we will compute the nonlinearity coefficients of an implanted or diffused resist. First, we will derive a relationship between the applied voltage over the resistor and the current through the resistor. We will assume that the voltage is the controlling quantity and the current the controlled one. Hence, we will represent the element as a conductance instead of a resistant

Figure 3.5 depicts an implanted or diffused resistor consisting of n-type material with a uniform doping concentration  $N_{res}$ . The applied voltage over the resistor is denoted as  $v_{CONTR}$ , a

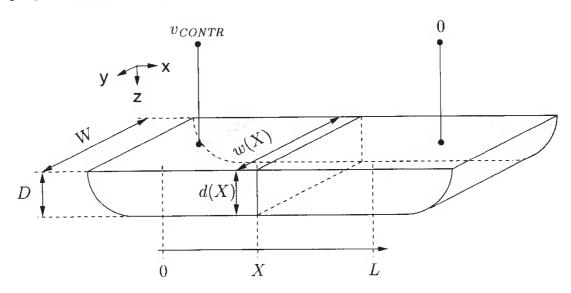


Figure 3.5: A diffused or implanted resistor.

the current through the element is written as  $i_{OUT}$ .

The resistor is embedded in uniform p-type material with a concentration  $N_p$ . The vertical distance between the surface and the junction is denoted as D. The horizontal width of resistance, from junction to junction is W. When a voltage  $v_{CONTR}$  is applied over the resistor current will flow through the resistor. It is assumed that this flow is only in the x-direction. The length of the resistor is L.

The effective depth and width of the resistor are reduced by the extension of the deplet layer inside the resistor. Since the width of a depletion layer depends on the voltage over pn-junction, the size of the resistor, and hence its value, is voltage-dependent. In the subsequence calculations, the nonlinearity caused by this effect will be computed.

In the calculations below, the effective depth and width will be denoted by d and w, respectively. These are given by

$$d = D - x_d \tag{3.88}$$

$$w = W - 2x_d \tag{3.89}$$

in which  $x_d$  denoted the extension of the depletion layer inside the n-type material. For an abrupt junction, this is given by [Sze 85]

$$x_d = A\sqrt{V_j + v} (3.90)$$

Here  $V_j$  is the junction potential and v is the voltage on an x-position between 0 and L, referred to the voltage on the place x = L. The factor A is given by

$$A = \sqrt{\frac{2\varepsilon_{Si}}{q} \left(\frac{1}{N_{res}} + \frac{1}{N_p}\right)} \cdot \frac{N_p}{N_p + N_{res}}$$
(3.91)

The junction potential  $V_j$  depends on the doping of the resistor  $N_{res}$  and of the surrounding p-type material,  $N_p$ :

$$V_j = \frac{kT}{q} \ln \frac{N_{res} N_p}{n_i^2} \tag{3.92}$$

It is assumed that the current in the resistor is totally determined by the majority carriers, in this case the electrons that drift due to the applied voltage. No recombination of generation of electron-hole pairs is considered. The current density of electrons that flow into the x-direction is then given by [Sze 85]

$$j_n(x) = -q\mu_n n \cdot \frac{dv(x)}{dx} = -\frac{1}{\rho_n} \cdot \frac{dv(x)}{dx}$$
(3.93)

in which n indicates the electron concentration per volume unit and  $\mu_n$  is the mobility of electrons;  $\rho_n$  is the bulk resistivity.

The total current as a function of x is found to be

$$i_{OUT}(x) = -\int_{y=0}^{y=w} \int_{z=0}^{z=d} \frac{1}{\rho_n} \cdot \frac{dv(x)}{dx}$$
 (3.94)

According to the above assumptions, the integrand is constant, such that

$$i_{OUT}(x) = -\frac{1}{\rho_n} \cdot w \cdot d \cdot \frac{dv(x)}{dx}$$
(3.95)

or

$$i_{OUT}(x) dx = -\frac{1}{\rho_n} \cdot w \cdot d \cdot dv(x)$$
(3.96)

Since recombination and generation of electron-hole pairs are neglected, the current can be considered as being constant between x=0 and x=L. Also, the bulk resistivity is assumed to constant in the x-direction. Integrating both sides of equation (3.95) from x=0 to x=L, as using equations (3.88) and (3.89) then yields

$$i_{OUT} = \frac{1}{\rho L} \int_0^{v_{CONTR}} \left( W - 2A\sqrt{V_j + v} \right) \left( D - A\sqrt{V_j + v} \right) dv \tag{3.9}$$

After some algebra one obtains

$$i_{OUT} = \frac{1}{\rho L} \left[ WD \cdot v_{CONTR} - \frac{2}{3} A \left( 2D + W \right) \left( \left( V_j + v_{CONTR} \right)^{3/2} - V_j^{3/2} \right) + 2A^2 V_j \cdot v_{CONTR} + A^2 v_{CONTR}^2 \right]$$
(3.9)

Equation (3.98) is an admittance description of the nonlinear resistor, expressing the current through the device as a function of the voltage over the device. This relationship is nonlinear, however, the extension of the depletion layer inside the resistor is neglected, which correspond to setting A=0, then the resistor becomes linear, as can be seen from the admittance description

$$i_{OUT} = \frac{WD}{\rho L} \cdot v_{CONTR} \tag{3.9}$$

For the determination of the nonlinearity coefficients equation (3.98) is developed into power series around the quiescent point  $(I_{OUT}, V_{CONTR})$ . This yields

$$i_{OUT} = I_{OUT} + \frac{1}{\rho L} \left[ \left( WD - \frac{A (2D + W) (V_j + V_{IN})^{3/2}}{V_j + V_{IN}} + 2A^2 (V_j + V_{CONTR}) \right) v_{contr} + \left( A^2 - \frac{A (2D + W)}{4\sqrt{V_j + V_{CONTR}}} \right) v_{contr}^2 + \left( \frac{A (2D + W)}{24 (V_j + V_{CONTR})^{3/2}} \right) v_{contr}^3 + \dots \right]$$
(3.16)

The AC part of this series can be identified with the power series expansion of the AC current through a nonlinear conductance (see Section 3.2.1):

$$i_{out} = g_{ac} \cdot v_{contr} + K_{2g_{ac}} \cdot v_{contr}^2 + K_{3g_{ac}} \cdot v_{contr}^3 + \dots$$

$$(3.10)$$

which yields

$$g_{ac} = \frac{1}{\rho L} \left( WD - A \left( 2D + W \right) \sqrt{V_j + V_{IN}} + 2A^2 \left( V_j + V_{CONTR} \right) \right)$$
 (3.10)

$$K_{2g_{ac}} = \frac{1}{\rho L} \left( A^2 - \frac{A(2D+W)}{4\sqrt{V_j + V_{CONTR}}} \right)$$
(3.10)

$$K_{3g_{ac}} = \frac{1}{\rho L} \left( \frac{A (2D + W)}{24 (V_j + V_{CONTR})^{3/2}} \right)$$
(3.16)

#### **3.3.1.1** Example

As an example, the nonlinearity coefficients as derived above are evaluated for an implanted resistor of  $20k\Omega$  which is realized with a phosphor or arsenic implantation in a p-type material. The depth D of the (n-type) implanted region is  $4.5\mu m$ . The impurity concentration  $N_n$  of the implanted region is  $1.1\times10^{16}cm^{-3}$ , whereas the concentration  $N_p$  of the underlying p-type material is  $8.1\times10^{14}cm^{-3}$ . The bulk resistivity  $\rho$  of the implanted region is  $0.4\Omega$  cm. The width W of the region is  $10\mu m$ . The junction potential (see equation (3.92)) is found to be 0.637V. The factor A of equation (3.91) is equal to  $9.08\times10^{-8}m/V^{1/2}$ .

If the extension of the depletion layer inside the n-type material is neglected, then the resistance is given by  $\rho L/(WD)$  (see the first term in the right-hand side of equation (3.98)). If this value is set equal to the required  $20k\Omega$ , then L is found to be  $225\mu m$ . We will now see what the influence of the extension of the depletion layer is, not only on the value of the resistor, but also on the nonlinearity coefficients.

Figure 3.6 shows the conductance of the resistor as a function of the voltage  $v_{CONTR}$  over the element. If the depletion layer would not extend at all into the implanted region, then the conductance would be  $5\times10^{-5}A/V$ . Due to this extension, the conductance is smaller as seen from Figure 3.6. Also, it is seen that the conductance decreases as  $v_{CONTR}$  increases. This is due to the increase of the extension of the depletion layer into the implanted region.

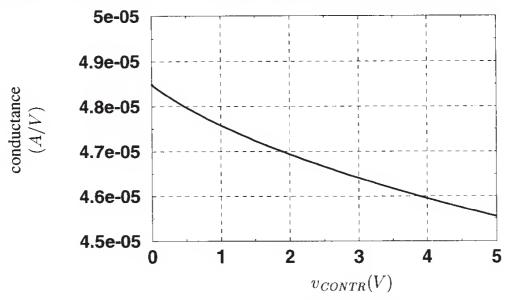


Figure 3.6: Variation of the conductance of an implanted resistor as a function of the voltage over the resistor.

In order to have an idea about the nonlinearity of the integrated resistor, the normalized nonlinearity coefficients of order two and three have been computed as a function of the input voltage as shown in Figure 3.7.

The voltage coefficient that expresses the nonlinearity of the AC conductance is found by multiplying  $K'_{2g_1}$  with two. In this way, we find from Figure 3.7 that the voltage coefficient of the resistor varies between 2.4% and 0.8%.

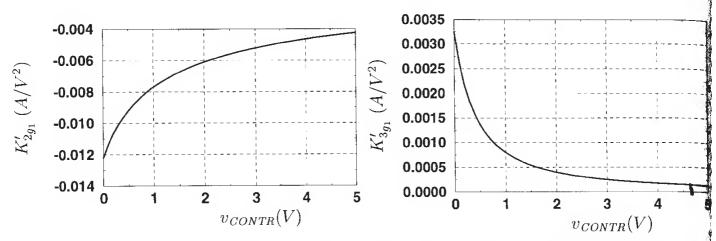


Figure 3.7: Normalized nonlinearity coefficient  $K'_{2g_1}$  (left) and  $K'_{3g_1}$  (right) for the implanted resistor as a function of the voltage over the resistor.

Further it is seen that both the second- and third-order nonlinearity coefficients becomes smaller in absolute value as  $v_{CONTR}$  increases. This means that the resistor becomes more linear as the voltage over the element increases.

#### 3.4 Integrated capacitors

The capacitors encountered in integrated circuits can roughly be divided in two types: oxide capacitors and junction capacitors. Oxide capacitors are consist of two conducting layers with an oxide in between. Examples are capacitors between a metal layer and a polysilicon layer, and capacitors between two different polysilicon layers. Junction capacitors, on the other hand, fall apart in two parts: the capacitor of the depletion layer around the junction, and the diffusion capacitor. The latter is only important when the junction is forwardly biased. In this section we concentrate on the nonlinearity of junction capacitors.

The junction capacitance per unit area is simply given by

$$C_j = \frac{\varepsilon_{Si}}{x_{dp} + x_{dn}} \tag{3.105}$$

in which  $x_{dp}$  and  $x_{dn}$  are the extension of the depletion layer into the p-type material and into the n-type material, respectively. For an abrupt junction,  $x_{dn}$  is given by equation (3.90), which is reformulated as

$$x_{dn} = \sqrt{\frac{2\varepsilon_{Si}}{q} \left(V_j + v_R\right) \left(\frac{1}{N_n} + \frac{1}{N_p}\right)} \cdot \frac{N_p}{N_p + N_n}$$
(3.106)

in which  $v_R$  is the reverse voltage over the junction,  $V_j$  the junction potential given by equation (3.92) and  $N_n$  and  $N_p$  are the impurity concentrations in the n-type material and the p-type material, respectively. The expression for  $x_{dp}$  is obtained by interchanging the role of p and n in equation (3.106).

Very often, the values  $N_p$  and  $N_n$  differ largely. If, for example,  $N_p \gg N_n$ , then  $x_{dn} \gg x_{dp}$ , and the junction capacitance per unit area can be approximated by

$$C_j \approx \sqrt{\frac{qN_n\varepsilon_{Si}}{V_j + v_R}} = \frac{C_{j0}}{\sqrt{1 + \frac{v_R}{V_j}}}$$
(3.107)

in which  $C_{j0}$  is the junction capacitor at zero bias:

$$C_{j0} = \frac{q\varepsilon_{Si}N_n}{\sqrt{2V_j}} \tag{3.108}$$

In many practical implementations a junction is not abrupt. In those situations, the junction capacitance per area unit is given by

$$C_{j} = \frac{C_{j0}}{\left(1 + \frac{v_{R}}{V_{i}}\right)^{m_{j}}} \tag{3.109}$$

The exponent  $m_j$  is referred to as the junction grading coefficient. For an abrupt junction,  $m_j = 0.5$  as explained above, and for a linear junction  $m_j = 0.33$  [Sze 85, Lak 94].

It is seen in equation (3.109) that  $C_j$  goes to infinity for  $v_R = -V_j$ . This does not occur in practice. In [Poon 69] it is shown that expression (3.107) is no longer valid at forward bias voltages around the junction potential. Nevertheless, the value of a junction capacitance at considerable forward bias voltages (say larger than  $V_j/2$ ) is not important, since under forward bias the junction capacitance is much smaller than the diffusion capacitance of the junction. As a result, distortion caused by junction capacitors under forward bias conditions is neglected compared to distortion caused by diffusion capacitors.

In Figure 3.8 the variation of the junction capacitor is shown as a function of the reverse voltage  $v_R$  over the junction for  $C_{j0}$  being 30fF,  $m_j=0.33$  and  $V_j=0.7V$ . The scale of the vertical axis is logarithmic. An interesting observation from Figure 3.8 is that at large reverse bias voltages the junction capacitance not only decreases, but it also changes less with the applied reverse voltage. Hence it becomes more linear at large reverse bias values.

For distortion computations it is interesting to derive the nonlinearity coefficients that describe a junction capacitance. To this purpose, equation (3.109) is expanded into a power series around the quiescent point. This yields

$$C_{j} = \frac{dQ}{dv_{R}} = C_{j}(V_{R}) + \frac{dC_{j}}{dv_{R}}v_{r} + \frac{1}{2} \cdot \frac{d^{2}C_{j}}{dv_{R}^{2}}v_{r}^{2} + \dots$$
(3.110)

in which Q is the charge associated with the depletion layer,  $V_R$  is the quiescent value of the reverse voltage  $v_R$  and  $v_r$  is the incremental value of the reverse voltage. The derivatives in equation (3.110) are evaluated at  $V_R$ .

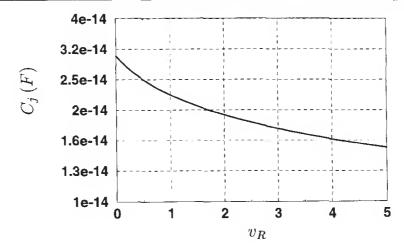


Figure 3.8: Variation of the junction capacitance as a function of the reverse voltage.

The AC current through the capacitor  $C_j$  is

$$i_{out} = \frac{dQ}{dt} = \frac{dQ}{dv_R} \cdot \frac{dv_R}{dv_r} \cdot \frac{dv_r}{dt} = C_j \frac{dv_r}{dt}$$
(3.111)

Substitution of equation (3.110) into equation (3.111) yields

$$i_{out} = C_j(V_R) \frac{dv_r}{dt} + \frac{1}{2} \cdot \frac{dC_j}{dv_R} \frac{d}{dt} \left(v_r^2\right) + \frac{1}{6} \cdot \frac{d^2 C_j}{dv_R^2} \frac{d}{dt} \left(v_r^3\right) + \dots$$
 (3.112)

This is identified with

$$i_{out} = C_j(V_R) \frac{dv_r}{dt} + K_{2C_j} \frac{d}{dt} (v_r^2) + K_{3C_j} \frac{d}{dt} (v_r^3) + \dots$$
 (3.113)

which yields

$$K_{2C_j} = \frac{1}{2} \cdot \frac{dC_j}{dv_R} \tag{3.114}$$

$$K_{3C_{j}} = \frac{1}{6} \cdot \frac{d^{2}C_{j}}{dv_{R}^{2}} \tag{3.115}$$

Using equation (3.109) one finds

$$K_{2C_{j}} = -\frac{1}{2} \cdot \frac{C_{j}}{1 + \frac{V_{R}}{V_{i}}} \cdot \frac{m_{j}}{V_{j}}$$
(3.116)

$$K_{3C_{j}} = \frac{1}{6} \cdot \frac{C_{j}}{\left(1 + \frac{V_{R}}{V_{j}}\right)^{2}} \cdot \frac{m_{j} \left(m_{j} + 1\right)}{V_{j}^{2}}$$
(3.117)

or, for the normalized nonlinearity coefficients

$$K'_{2C_j} = -\frac{1}{2} \cdot \frac{1}{1 + \frac{V_R}{V_i}} \cdot \frac{m_j}{V_j}$$
(3.118)

$$K'_{3C_j} = \frac{1}{6} \cdot \frac{1}{\left(1 + \frac{V_R}{V_j}\right)^2} \cdot \frac{m_j (m_j + 1)}{V_j^2}$$
(3.119)

It is seen that the normalized nonlinearity coefficients decrease when the reverse voltage increases. This is in correspondence with the observation made from Figure 3.8: a junction capacitor becomes more linear as the reverse voltage increases.

As an example, consider a base-collector junction with a junction potential  $V_j$  of 0.7V and a grading coefficient  $m_j$  of 0.33. Then one obtains  $K'_{2C_j} = -0.23V^{-1}$  and  $K'_{3C_j} = 0.15V^{-2}$  for a reverse voltage of 0V. For a reverse voltage of 1V one obtains  $K'_{2C_j} = -0.09V^{-1}$  and  $K'_{3C_j} = 0.025V^{-2}$ .

Compared to other integrated capacitors, a junction capacitor is quite nonlinear. This is due to the change of the width of the depletion layer with the applied voltage over the junction. For example, the voltage coefficient of an integrated poly-substrate capacitor is 0.05%/V, and for a poly-poly capacitor 0.005%/V [Lak 94].

#### 3.5 Weakly nonlinear transistor models: introduction

The different basic nonlinearities described in the previous sections can now be tailored together to construct nonlinear equivalent circuits for transistors. These are straightforward extensions of the linear equivalent circuits [Gray 93, Lak 94]. Before discussing silicon bipolar and MOS models in depth in Chapters 6 and 7, a few preliminary concepts are discussed here.

Much like many transistor models that are used in SPICE-like circuit simulators, the transistor models that will be used in this book contain some simplifications. In this way, some insight can be obtained in the nonlinear phenomena that are the major sources of distortion caused by a transistor. More accurate models are sometimes too complicated to provide such insights and require iteration or finite-element analysis for their evaluation. A considerable simplification both for bipolar and MOS models is achieved by considering the current flow in a transistor as one-dimensional. Explicit two-dimensional effects are then modeled as extensions to those one-dimensional models. Also, distributed elements are lumped into just a few circuit elements. An example is the base resistance of a bipolar transistor which is shown in Figure 3.9.

It is seen that the base resistance is distributed over the width of the base-emitter junction. Such distributed representation is not practical to handle in circuit design. Instead, a lumped representation is used.

Apart from the simplifications as the ones mentioned above, some further simplifications are performed in practical device models, still in order to end up with tractable analytic expressions.

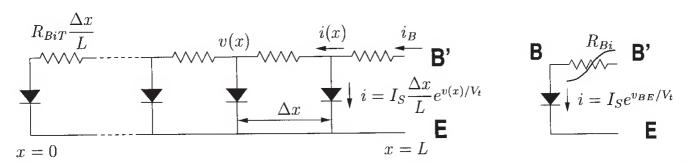


Figure 3.9: The intrinsic base resistance and the base-emitter junction of a bipolar transistor consist of distributed elements (left). In bipolar transistor models for circuit design these elements are lumped into only two circuit elements (right). The extrinsic base resistance (not shown) is put in series with  $R_{Bi}$ .

For the analysis of a linearized circuit first-order derivatives of the model equations are required and — as pointed out in this chapter — for distortion computations higher-order derivative are required. For example, the transconductance  $g_m$  of a MOS transistor is obtained by taking the derivative of the model equation for the drain current with respect to  $v_{GS}$ . The second and third-order nonlinearity coefficients  $K_{2g_m}$  and  $K_{3g_m}$  are — according to the results from Section 3.2.1 — obtained by taking the second- and third-order derivative of the drain current However, some effects which do not seem to be important for the computation of the current of  $g_m$  could become important when the higher-order derivatives are considered. Hence, one should be careful not to oversimplify the models and to introduce fit factors too rigorously. However, an accurate computation of derivatives of order five or higher is not achievable due to the inevitable simplifications that have been made to obtain analytic model equations.

Existing circuit simulators provide values for the transistor current and its first derivatives which are the small-signal parameters, but the higher-order derivatives — which are proportional to the nonlinearity coefficients — are not computed. In order to obtain a value for the nonlinearity coefficients, two approaches can be followed. First, the derivatives can be computed by numerical differentiation [Num 92]. This approach can lead to inaccuracies, since numerical differentiation is a process that can lead to numerical instabilities. Furthermore, this approach does not yield enough insight: apart from the numerical values, it is not known which physical effects determine the nonlinearity coefficient under consideration.

In order to avoid the disadvantages of numerical differentiation, an alternative approach has been developed. In this approach, analytical expressions for the derivatives are determined by computing symbolically the derivatives of the model equations with respect to the different controlling voltages or currents. This process can be speeded up by making use of symbolic algebra programs such as MAPLE [Map 91], MACSYMA [Macs 87] or MATHEMATICA [Wolf 91], that can compute derivatives symbolically. These programs provide facilities to dump a given expression into source code for C or Fortran. Using this approach, C routines have been developed. In these routines the different derivatives are computed as a function of the bias voltages or currents, the transistor dimensions and the model parameters. The correctness of the routines

has been checked by a comparison with the derivatives that have been computed by numerical differentiation.

An additional test for the correctness is a comparison with simulation results from a numerical circuit simulator. As already mentioned above, these circuit simulators usually do not compute derivatives of order higher than one. Nevertheless, a comparison of the current and the first derivatives with simulation results already give a good indication of the correctness. This comparison, of course, is only possible for models that have also been implemented in the circuit simulator.

The expressions of derivatives of model equations are usually very complicated. Insight can be obtained from those expressions by plotting them as a function of bias, transistor dimensions, model parameters, .... Nevertheless, the insight in distortion phenomena would be greatly enhanced if a closed-form expression — even if it is approximate — could be obtained for the derivatives of interest. To this purpose, the C routines mentioned above have been extended by routines that trace the dominant contributions to a given nonlinearity coefficient. An interpretation of these dominant contributions makes it possible to identify the dominant physical effects that determine a given nonlinearity coefficient.

A tracing of the dominant terms of a derivative is possible since the transistor current is usually a sum or a product of several composed functions. For example, in Chapter 7 it will be shown that the drain current of a MOS transistor in strong inversion (triode region) can be written as a product of three functions:

$$i_D = mobred(v_{GS}, v_{DS}, v_{SB}) \cdot hot(v_{GS}, v_{DS}, v_{SB}) \cdot large(v_{GS}, v_{DS}, v_{SB})$$
 (3.120)

in which the three functions mobred, hot and large are in turn functions that may contain sums, products, powers and exponential functions of the transistor terminal voltages  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ . The meaning of these three functions will be explained in Chapter 7. Since the derivative of a product with respect to a variable (in this case a voltage) is a sum of two functions, it is easy to see that the nonlinearity coefficients, which are proportional to higher-order derivatives of the current with respect to the terminal voltages, will consist of a sum with many terms. An identification of the dominant terms in this sum will yield an approximate, interpretable expression of the nonlinearity coefficient under consideration. This approach will be illustrated in Section 7.7.4.

The C routines that are developed to evaluate the nonlinearity coefficients and to identify the dominant contributions could be combined with a numerical circuit simulator. Moreover, it is not mandatory that the transistor model used by the circuit simulator is exactly the same as the model that is used in the computations of the nonlinearity coefficients: the circuit simulator only needs to find the DC solution, providing the bias values needed to evaluate the nonlinearity coefficients.

### 3.6 Summary

In this chapter we defined the nonlinearity coefficients of the basic nonlinearities that will be used in the next sections to describe the nonlinear behavior of bipolar and MOS transistors. In addition, we discussed the nonlinearity of integrated capacitors and resistors.

The nonlinearity coefficients of order two and three are proportional to second- and third-order derivatives of the model equation with respect to the controlling voltage(s) or current(s).

Since model equations can already be quite complicated, the expressions of the derivatives are even more complicated such that they cannot be interpreted easily by a design engineer. In order to overcome this problem, an approach has been discussed to derive the dominant terms of the nonlinearity coefficients. This results in shorter expressions which can be better interpreted. This approach will be used in Chapters 6 and 7.

í

## **Chapter 4**

# Volterra series and their applications to analog integrated circuit design

#### 4.1 Introduction

In this chapter it is investigated how weakly nonlinear behavior of an analog integrated circuit can be computed. The emphasis here is on obtaining insight. Therefore, less attention is paid to numerical methods such as time-domain analysis followed by a Fourier transform, which is the classical SPICE approach [Royc 89, Hspi 96], or harmonic balance methods. Such methods compute the response of a nonlinear circuit by iteration, and the final result is a list of numbers, which do not indicate which nonlinearities in the circuit are mainly responsible for the observed nonlinear behavior. Hence such methods are suitable for verification of circuits that have already been designed. When simulations show that the specifications regarding the nonlinear behavior are not met, then these methods do not present information from which designers can derive which circuit parameters or circuit elements they have to modify in order to obtain the required specifications. Such valuable information can be obtained with the use of Volterra series. The price that is paid for this extra insight is that the analysis is limited to weakly nonlinear behavior only.

Volterra series have already been used for distortion computations [Nar 67, Nar 70, Mey 72, Sans 72, Kuo 73, Nar 73, Buss 74, Khad 74, Kuo 77, Rud 78, Chua 79a, Chua 79b, Wein 80, Sale 82, Chua 82, Maas 88, Wamb 90], also with SPICE [Chis 73, Royc 89]. In several implementations of SPICE these computations can be performed with the .DISTO command. Nevertheless, Volterra series are seldom used by IC designers. This is explained by the fact that the use of Volterra series is limited to weakly nonlinear behavior, and the simulation results are presented in the same way as for most numerical simulators, namely as a list of numbers, from which it is not clear what is behind those numbers. In this book we try to perform distortion computations with Volterra series in a way that insight can be gained.

Volterra series describe the output of a nonlinear system as the sum of the response of a first-order operator, a second-order one, a third-order one and so on [Sche 80]. These operators are shown in the block-diagram representation of Figure 4.1. Every operator is described either in

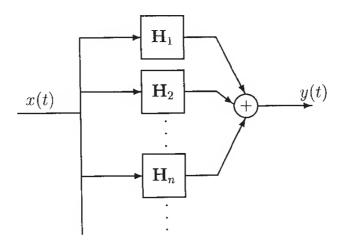


Figure 4.1: Schematic representation of a system characterized by a Volterra series.

the time domain or in the frequency domain with a kind of transfer function, called a *Volterra kernel*. Just as for linear circuits, the frequency-domain representation is preferred over the time-domain representation for many circuit analysis aspects. It has to be remarked here that the block that is represented in Figure 4.1 with  $\mathbf{H}_1$  represents the linearized system.

In fact a Volterra series describes a nonlinear system in a way which is equivalent to the way Taylor series approximate an analytic function. A nonlinear system which is excited by a signal with small amplitude can be described by a Volterra series which can be broken down after the first few terms. The higher the input amplitude, the more terms of that series need to be taken into account in order to describe the system behavior properly. For very high amplitudes, the series diverges, just as Taylor series. This is often due to the occurrence of a strong nonlinearity at high amplitude levels, such as the cut-off of a transistor. Hence, Volterra series are only suitable for the analysis of weakly nonlinear circuits.

The Volterra series approach has been proven to be very attractive for hand calculations of small transistor networks. Several studies of effects like intermodulation or harmonic distortion and cross modulation in such circuits have been reported [Nar 67, Mey 72, Poon 72, Sans 72, Nar 73, Khad 74, Abra 76]. Since Volterra kernels retain phase information, they are especially useful for high-frequency analysis and for effects like AM to PM conversion. Volterra series give a general characterization of a nonlinear circuit in the sense that once the Volterra kernels of a circuit is known, its output can be found for any input. For example, the response of a nonlinear system to noise can be studied with Volterra series [Bedr 71, Rud 78, Sche 80]. In Section 4.2 some definitions concerning Volterra series are given. The mathematical foundations of those definitions are given in Appendix B.

The Volterra series representation is not only an explicit nonlinear representation of the system response in terms of the input, but also provides insight into the system operation. This insight is readily obtained from the block-diagram representation of the Volterra kernels. Using this representation some interesting applications can be considered like cascading of nonlinear circuits, distortion cancellation, linear and nonlinear feedback in weakly nonlinear circuits,....

These topics are covered in the Sections 4.5, 4.6, 4.7 and 4.8.

The reported hand calculations of Volterra kernels were limited to small circuits only. In the seventies an algorithm has been developed for the numerical calculation of Volterra kernels in the frequency domain for larger nonlinear networks. The development of the original algorithm by Bussgang, Ehrman and Graham [Buss 74] was granted by the U.S. Air Forces for the analysis of their communication receivers. Later, this algorithm has been generalized by Chua and Ng [Chua 79b]. This algorithm is explained in Chapter 5. It is interesting to note that a similar approach for time-discrete circuits such as switched-capacitor circuits, has been described in [VdWal 83]. In Chapter 5 some alternative methods will be discussed that circumvent the use of Volterra series, but the approach is basically the same.

The Volterra series approach is not the only technique to obtain approximate closed-form expressions for nonlinear behavior. Another technique is the method of the describing function [Ath 75]. With this technique it is possible to obtain closed-form expressions for a feedback system that contains an isolated static nonlinearity in the feedback loop. Since it is not possible in general to map an analog integrated circuit to such a feedback system, the method of the describing function is not used here.

#### 4.2 Basics of Volterra series

In this section the fundamentals of Volterra series are briefly discussed, both in the time domain and in the frequency domain. A thorough study of Volterra series can be found in [Sche 80].

The theory of Volterra series can be viewed as an extension of the theory of linear, first-order systems to weakly nonlinear systems [Sche 80]. In the Volterra series description, such a system is considered as the combination of different operators of different order as shown in Figure 4.1. Every block  $\mathbf{H}_1$ ,  $\mathbf{H}_2$  and  $\mathbf{H}_n$  represents an operator of order  $1, 2, \ldots$ , respectively.

When a nonlinear system is excited by a signal with very low input amplitude, then the output can be described accurately by taking into account only the first-order behavior of the system, which is the linear behavior, represented by block  $H_1$ . In the frequency domain this block is characterized by the transfer function of the linearized circuit. When the input amplitude increases, then a substantial part of the output signal is caused by nonlinear effects. For a sufficiently low input amplitude, these nonlinear effects can be described accurately by taking into account second- and third-order effects only, which are modeled by the operators  $H_2$  and  $H_3$  of Figure 4.1. The total output is the sum of the outputs of every operator individually. When the input amplitude increases even more, then higher-order effects occur, which are modeled by higher-order operators.

For a nonlinear system excited by a sine wave with a limited amplitude this model roughly corresponds to a decomposition of the output signal into the different harmonics. The nth harmonic is primarily determined by the nth-order operator of Figure 4.1. Just as for linear circuits, the analysis can be performed in the time domain as well as in the frequency domain.

The use of Volterra series for nonlinear systems can be compared to the use of Taylor series for analytic functions. In the latter case, small excursions of the function arguments around a fixed point can be described accurately, if at least the excursions fall within the convergence

radius of the Taylor series. The larger the excursions are, the more terms of the Taylor series must be taken into account for a sufficiently accurate approximation. Also, some Taylor series around a certain point have a convergence radius of zero. Equivalently, Volterra series have a convergence radius which can be zero. The latter case corresponds to systems which cannot be described at all with a converging Volterra series such as an ideal clipper [Sche 80] or an idea comparator. Figure 4.2 shows the input-output relationship of such clipper. When the input

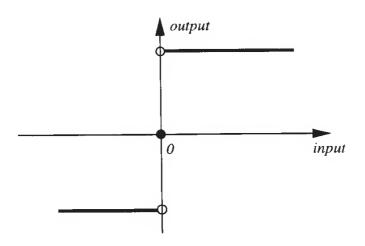


Figure 4.2: Input-output relationship of an ideal clipper.

signal is zero, then the output is zero as well. The least deviation of the input signal from zero i either direction gives an output of +1 or -1. It is clear that this is a strongly nonlinear effect.

#### 4.2.1 Volterra operators

Consider a second-order nonlinear system or, briefly, a second-order nonlinearity. Its operatic can be seen as follows. The second-order nonlinearity combines two signals, which eventuall are identical, and produces with these signals a second-order signal. A third-order nonlinearit combines three signals, which can be identical, then multiplies them to produce a third-order signal. The combination of signals can be just a multiplication, but in Sections 4.3.1 and 4.3. we will see that the two signals can be combined in a more complex way.

Assume now that a single sinusoidal signal with a frequency  $\omega_1$  is applied to a second-order nonlinearity. In this case the second-order nonlinearity combines two times this signal to product a second-order signal. This signal has two frequency components, one at  $\omega_1 + \omega_1 = 2\omega_1$  are one at  $\omega_1 - \omega_1 = 0$  Hz. When two sinusoidal signals with frequency  $\omega_1$  and  $\omega_2$  are applied this second-order nonlinearity, then the second-order signal combines two times the first sign to produce a DC term and a component at  $2\omega_1$ , which is denoted as a second harmonic. Also the second-order system combines two times the input signal at  $\omega_2$  to produce a DC term and component at  $2\omega_2$ . Finally, the nonlinearity also combines the sine wave at  $\omega_1$  with the sine way at  $\omega_2$  to produce a sine wave at the frequencies  $\omega_1 + \omega_2$  and  $|\omega_1 - \omega_2|$ . The latter two response are intermodulation products.

Consider now a third-order nonlinearity. This nonlinearity will combine three signals, some of which can be identical, and produce with these signals a third-order signal. When one sinusoidal signal at frequency  $\omega_1$  is applied to this nonlinearity, then the latter produces sine waves at the frequencies  $|\pm \omega_1 \pm \omega_1|$ . This corresponds to two frequencies, namely  $\omega_1$  and  $3\omega_1$ .

With the above considerations we obtain components at the same frequencies we found in Chapter 2 for a simple nonlinear system. Although the considerations look quite intuitive, they can be formalized using second-order, third-order and in general *n*th-order operators. This is explained in Appendix B.

An operator performs a transformation of the input signal that results in an output signal. The operator that corresponds to the second-order nonlinearity that we discussed above, is denoted here as the **second-order Volterra operator** and its notation is  $\mathbf{H}_2$ . The output of this second-order operator to which an input signal x(t) — which can be a sine wave, a combination of sine waves, . . . — is applied, is then denoted as  $\mathbf{H}_2[x(t)]$ . Similarly, an *n*th-order nonlinearity is described with an *n*th-order Volterra operator  $\mathbf{H}_n$ , and the output of this *n*th-order nonlinearity is denoted as  $\mathbf{H}_n[x(t)]$ . Finally, one should not forget to model the linear behavior of the system as well. This is performed with a **first-order Volterra operator**. This operator is identical to the operator that describes a linear system, and its computation can be performed using the theory of linear systems. From this theory we know that the transformation that is performed by this first-order Volterra operator is a convolution of the input signal with the impulse response of the linear or linearized system.

#### 4.2.2 Time-domain fundamentals

With the Volterra series model of Figure 4.1 a nonlinear system consists of several Volterra operators in parallel. In the previous section we explained qualitatively how such operators act on an input signal. In this section we will explain this in a quantitative manner.

In Appendix B it is shown that the transformation on an input signal performed by a secondorder Volterra operator is given by

$$\mathbf{H}_{2}[x(t)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) x(t - \tau_{1}) x(t - \tau_{2}) d\tau_{1} d\tau_{2}$$
(4.1)

This two-dimensional integral is recognized as a two-dimensional convolution integral. Here we see a similarity with a linear system: whereas the output of a linear system is computed by taking the convolution of the input signal with the impulse response of the linear system, the output of a second-order nonlinearity is a two-dimensional convolution of the function  $h_2(\tau_1, \tau_2)$  with the input signal. The function  $h_2(\tau_1, \tau_2)$  is denoted as the **second-order Volterra kernel**. In [Sche 80] it is shown that the second-order Volterra kernel can be considered as a two-dimensional impulse response.

Consider now a third-order Volterra operator. The mathematical representation of the transformation that is performed by a third-order Volterra operator, is given by

$$\mathbf{H}_{3}[x(t)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{3}(\tau_{1}, \tau_{2}, \tau_{3}) x(t - \tau_{1}) x(t - \tau_{2}) x(t - \tau_{3}) d\tau_{1} d\tau_{2} d\tau_{3}$$
 (4.2)

and, for an nth-order Volterra operator

$$\mathbf{H}_{n}[x(t)] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} h_{n}(\tau_{1}, \tau_{2}, \dots, \tau_{n}) x(t - \tau_{1}) x(t - \tau_{2}) \cdots x(t - \tau_{n}) d\tau_{1} d\tau_{2} \dots d\tau_{n}$$

$$\tag{4.3}$$

The *n*-dimensional integral is seen to be an *n*th-order convolution integral. The function  $h_n(\tau_1, \tau_2, \dots, \tau_n)$  is an *n*th-order Volterra kernel. In [Sche 80] it is proven that this can be considered as an *n*th-order impulse response. Equation (4.3) indicates that the output signal of an *n*th-order Volterra operator can be computed by taking the *n*-dimensional convolution of this *n*-dimensional impulse response with the input signal.

Recall now that with the formalism of the Volterra series the output of a nonlinear system is represented as the sum of the output of a first-order Volterra operator with the output of a second-order one, a third-order one, and so on, as shown in Figure 4.1. Using equations (4.1), (4.2) and (4.3) we find that the complete output of the complete nonlinear system is given by

$$y(t) = \mathbf{H}_1[x(t)] + \mathbf{H}_2[x(t)] + \mathbf{H}_3[x(t)] + \dots + \mathbf{H}_n[x(t)] + \dots$$
(4.4)

This equation is a Volterra series representation of the nonlinear system.

Let us consider the *n*th-order Volterra kernel  $h_n(\tau_1, \tau_2, \dots, \tau_n)$  in more detail. It is seen that a Volterra kernel of order n is a function of n time-domain variables. In Appendix B it is shown that a kernel can always be made *symmetric*, which means that the value of the kernel does not change when the arguments are interchanged. All Volterra kernels that will be used throughout this book are assumed to be symmetric unless it is explicitly mentioned they are not symmetric.

In this book we are reasoning only with causal systems. This means that the response does not depend on the future of the input. For such systems it can be proven [Sche 80] that the value of the Volterra kernels for any negative argument is zero:

$$h_n(\tau_1, \tau_2, ..., \tau_n) = 0$$
 for any  $\tau_j < 0$ ,  $j = 1, 2, ..., n$  (4.5)

Let us now consider the shape of a Volterra series. The Volterra series is a power series. This can be seen by changing the input by a factor a, resulting in a new input ax(t). Using equation (4.3) and (4.4) the new output is

$$y(t) = \sum_{n=1}^{\infty} \mathbf{H}_n[ax(t)]$$
$$= \sum_{n=1}^{\infty} a^n \mathbf{H}_n[x(t)]$$
(4.6)

Moreover, a Volterra series is a series with memory. Indeed, the integrals for  $\mathbf{H}_n$  are convolutions. This is obvious for the first term of the series in equation (4.4), which is nothing else but the convolution integral of a linear system. The terms after the first one are higher-order convolutions [Sche 80]. This means that Volterra series —in contrast with power series—can

describe weakly nonlinear circuits with capacitors and inductors. If the system described by the Volterra series of equation (4.4), has no memory at all, then

$$h_n(\tau_1, \tau_2, ..., \tau_n) = 0$$
 for any  $\tau_j > 0$ ,  $j = 1, 2, ..., n$  (4.7)

and the Volterra series reduces to the power series

$$y(t) = \sum_{n=1}^{+\infty} h_n(0, 0, \dots, 0) x^n(t)$$
 (4.8)

As a result of its power series character, there are some limitations associated with the use of Volterra series to nonlinear problems. Just as with the Taylor series representation of a function, problems of convergence may occur. When the input amplitude increases, the higher-order terms in the Volterra series relatively increase more than the lower-order ones. At a certain amplitude, divergence may occur. Hence, a convergence radius can be defined [Boyd 84]. The computation of a convergence radius is quite complicated and it will not be considered in this book. Instead it will be assumed that the signals that are applied to the circuits under consideration are small enough such that the Volterra series converges. This is a realistic assumption in most situations.

#### 4.2.3 Frequency-domain representation

From the theory of linear systems we know that the output of a linear system can be computed in the time domain by performing a (one-dimensional) convolution of the time-domain representation of the input signal with the system's impulse response. In the frequency domain the response of a linear system can be computed by multiplying the Fourier transform of the impulse response with the Fourier transform of the input signal. This approach can now be extended to second-order, third-order, ... operators. It has already been mentioned above that an *n*th-order Volterra kernel can be considered as an *n*th-order impulse response. Consequently, if we draw the parallel with a linear system, we expect that the Fourier transform of the output of an *n*th-order Volterra operator involves a multiplication with the Fourier transform of the *n*th-order Volterra kernel. This is considered now in more detail.

First we consider the frequency domain representation of an nth-order Volterra kernel. To this purpose a multidimensional Laplace and Fourier transform are defined. The Laplace transform of the nth-order Volterra kernel kernel is called the nth-order nonlinear transfer function or the nth-order kernel transform. The analogy with a linear transfer function is again due to the fact that an nth-order Volterra kernel can be considered as an n-dimensional impulse response. Very often, when no confusion is possible, the frequency domain representation of the kernel is also called a Volterra kernel. Since the nth-order kernel is a function of n time variables, a complete frequency domain representation requires n frequency variables.

In Appendix B it is shown that the *multidimensional Laplace transform*  $H_n(s_1, \ldots, s_n)$  of the kernel  $h_n(\tau_1, \ldots, \tau_n)$  is given by

$$H_n(s_1,\ldots,s_n) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} h_n(\tau_1,\ldots,\tau_n) e^{-(s_1\tau_1+\ldots s_n\tau_n)} d\tau_1 \cdots d\tau_n$$
 (4.9)

in which  $s_i = \sigma_i + j\omega_i$  (i = 1, 2, ..., n) is a complex number. The **multidimensional Fourier** transform is obtained from the Laplace transform by making all  $\sigma_i$  zero in equation (4.9)<sup>1</sup>.

It can be proven easily that the nth-order transfer function is symmetric when its time-domain equivalent is symmetric. This is in fact very natural: a nonlinear operator cannot distinguish the different applied frequencies. An interesting property for symmetric nth-order transfer functions is that its complex conjugate can be obtained by changing the sign of the frequency arguments:

$$H_n(-j\omega_1, -j\omega_2, \dots, -j\omega_n) = H_n^*(j\omega_1, j\omega_2, \dots, j\omega_n)$$
(4.10)

Equivalent to the linear transfer function, the *n*th-order nonlinear transfer function can be used to find the *n*th-order system's output as a result of a sinusoidal excitation. Consider for example a second-order system that corresponds to the operator  $H_2$  in Figure 4.1. In Appendix B it is proven that the output of such a system excited by a sinusoidal excitation  $A_x \cos \omega_x t$  can be written in terms of the second-order nonlinear transfer function  $H_2(s_1, s_2)$  as follows:

$$y_2(t) = \frac{A_x^2}{2} \operatorname{Re} \left( H_2(j\omega_x, j\omega_x) e^{j2\omega_x t} \right) + \frac{A_x^2}{2} \operatorname{Re} \left( H_2(j\omega_x, -j\omega_x) \right)$$

$$= \frac{A_x^2}{2} |H_2(j\omega_x, j\omega_x)| \cdot \cos \left( 2\omega_x t + \arg \left( H_2(j\omega_x, j\omega_x) \right) \right) + \frac{A_x^2}{2} H_2(j\omega_x, -j\omega_x) \quad (4.11)$$

Clearly, the output of this second-order system is a constant and a sinusoid at frequency  $2\omega_x$ . These frequency components were found as well in Chapter 2 for a memoryless second-order nonlinearity. An analysis of the response to a sum of sinusoidal signals is postponed to Section 4.4.

For a nonlinear system corresponding to a third-order Volterra operator we find in Appendix B that the output to the sinusoidal excitation  $A_x \cos(\omega_x t)$  in terms of the third-order nonlinear transfer function  $H_3(s_1, s_2, s_3)$  is equal to

$$y_3(t) = \frac{A_x^3}{4} \operatorname{Re} \left( H_3(j\omega_x, j\omega_x, j\omega_x) e^{j3\omega_x t} \right) + \frac{3A_x^3}{4} \operatorname{Re} \left( H_3(j\omega_x, j\omega_x, -j\omega_x) e^{j\omega_x t} \right)$$

$$= \frac{A_x^3}{4} |H_3(j\omega_x, j\omega_x, j\omega_x)| \cdot \cos \left( 3\omega_x t + \arg(H_3(j\omega_x, j\omega_x, j\omega_x, j\omega_x)) \right)$$

$$+ \frac{3A_x^3}{4} |H_3(j\omega_x, j\omega_x, -j\omega_x)| \cdot \cos \left( \omega_x t + \arg(H_3(j\omega_x, j\omega_x, -j\omega_x)) \right)$$

$$(4.12)$$

It is seen that the response consist of a component at the frequencies  $3\omega_x$  and  $\omega_x$ . These frequencies have been found as well in Chapter 2 for a memoryless third-order nonlinearity. The difference between the results obtained with Volterra series and the results from Chapter 2 is that the result given here is more general: the Fourier transforms  $H_2(j\omega_1,j\omega_2)$  and  $H_3(j\omega_1,j\omega_2,j\omega_3)$  are complex numbers that model phase shifts on the harmonics due to capacitors or inductors in the nonlinear network. This was not taken into account in the simple calculations of Chapter 2, where we only calculated with real numbers instead of complex numbers. For nonlinear circuits where capacitive or inductive effects can be neglected, for example in amplifiers at low frequencies, it is clear that the responses computed in Chapter 2 must be the same as the responses

<sup>&</sup>lt;sup>1</sup>This, of course, requires that the region of absolute convergence includes the  $\omega_1, \omega_2, \ldots$  and  $\omega_n$  axes.

expressed in terms of Volterra series. This will be discussed further in Sections 4.3.1 and 4.3.2. Further, the computation of the Fourier transforms  $H_2(j\omega_1,j\omega_2)$  and  $H_3(j\omega_1,j\omega_2,j\omega_3)$  in a non-linear circuit is explained in Chapter 5.

For the sake of completeness we repeat that the response of a linear or first-order system to a sinusoidal excitation  $A_x \cos \omega_x t$  is given by

$$y_1(t) = A \operatorname{Re} \left( H_1(j\omega_x) e^{j\omega_x t} \right)$$
  
=  $A |H_1(j\omega_x)| \cos \left( \omega_x t + \arg(H_1(j\omega_x)) \right)$  (4.13)

#### 4.2.4 Weakly nonlinear circuit behavior revisited

At this point, weakly nonlinear circuit behavior can be defined in terms of Volterra series. The most general definition would be that a circuit excited by a given source behaves weakly nonlinearly if its response can be represented by a converging Volterra series. Although in some cases many terms of the Volterra series are required to accurately describe the circuit response even at moderate input levels [VdEi 89], the response of many circuits can most often be described by the first few terms of the Volterra series that describes the circuit under consideration. Therefore, the following more restrictive definition is more practical:

A circuit behaves weakly nonlinearly if, for the applied input signal, it can be accurately described by the first three terms of its (converging) Volterra series.

This definition requires that for the characterization of a circuit its linear behavior be computed together with its lowest even- and odd-order nonlinear behavior. This yields a quite complete description of the behavior of a large class of analog integrated circuits under normal signal conditions. One could argue that the above definition is mathematically not exact: it is not described how accurately the response of the circuit must be described. The required accuracy depends on the requirements of the circuit engineer. Most often a fairly low accuracy (for example errors up to a few dB) are acceptable if at least the results obtained using Volterra series yield insight in a circuit's nonlinear behavior.

For both the general and the more practical definition, the problem remains that it is in general not known in advance if the Volterra series will converge, since the radius of convergence can only be computed once the Volterra kernels are known. If the series converges, then it is generally not known how many terms are required for an accurate description [VdEi 89]. For the circuits to which the computations in the subsequent chapters are addressed, however, this is seldom a problem, as will be demonstrated with numerical simulations with other techniques that are not limited to weakly nonlinear behavior. Hence, the practical but more restrictive definition of nonlinear behavior can be used for most analog integrated circuits instead of the more general definition.

Finally, it must be noted that strictly speaking, weakly nonlinear behavior is not a property of the circuit alone, since it depends also on the amplitude of the excitation(s). However, many circuits are called weakly nonlinear or quasi-linear [Chua 75]. This means that they behave weakly nonlinearly when they are excited by practical signal levels for which they have been designed.

#### 4.3 Examples of Volterra kernels

The mathematical treatment of Volterra series in the previous sections is now clarified with a few examples.

#### 4.3.1 Basic second-order system

In this section we concentrate on a general block diagram representation of a second-order Volterra operator, which corresponds to the block  $H_2$  in Figure 4.1. It is clear that a second order operator performs at least one multiplication, since we stated that a second-order nonlinearity combines two signals to produce a second-order signal. The most general second-order system with only one multiplication is shown in Figure 4.3.

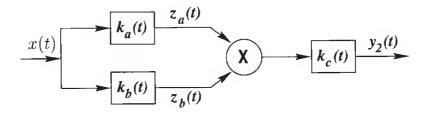


Figure 4.3: Block-diagram representation that illustrates the operation of a simple second-ord system.

With this simple second-order system the operation of a second-order nonlinearity can explained generally as follows: the incoming signal x(t) is first fed to two linear blocks wi impulse response  $k_a(t)$  and  $k_b(t)$ , respectively, yielding the outputs  $z_a(t)$  and  $z_b(t)$ . The latt signals are now combined by the multiplier block that produces a second-order signal, which turn is fed to a linear system characterized by the impulse response  $k_c(t)$ . The output of the linear system is the overall output  $y_2(t)$ .

The reader should be aware that the representation of Figure 4.3 is a general block-diagram representation in which the different sub-blocks do not necessarily correspond to "physical" subblocks in a practical nonlinear circuit. The block diagram has merely been used to explain how in general a second-order nonlinearity operates.

It is not difficult to prove [Sche 80] that the (second-order) Volterra kernel of this general system is given by

$$h_2(\tau_1, \tau_2) = \int_{-\infty}^{+\infty} k_c(\sigma) k_a (\tau_1 - \sigma) k_b (\tau_2 - \sigma) d\sigma$$
(4.14)

In this book we are more interested in a frequency-domain representation where we make us of Fourier transforms. The Fourier transform of  $h_2(\tau_1, \tau_2)$ , denoted as  $H_2(j\omega_1, j\omega_2)$  is found after some algebra — to be

$$H_2(j\omega_1, j\omega_2) = K_a(j\omega_1)K_b(j\omega_2)K_c(j\omega_1 + j\omega_2)$$
(4.1)

where  $K_a(j\omega)$ ,  $K_b(j\omega)$  and  $K_c(j\omega)$  denote the Fourier transforms of the linear subsystems described by the impulse responses  $k_a(t)$ ,  $k_b(t)$  and  $k_c(t)$ , respectively.

Let us now consider some interesting simplifications of the general block diagram of Figure 4.3. Assume that the the systems described by  $k_a(t)$ ,  $k_b(t)$  are not present, while  $k_c(t)$  corresponds to a multiplication with a constant scale factor  $K_2$ . This leads to the simplified block diagram of Figure 4.4. This corresponds to a second-order nonlinearity without memory, the same as we used in Chapter 2 in our simple calculations.

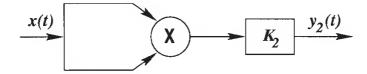


Figure 4.4: Block-diagram representation of a simple memoryless second-order system.

Comparing Figure 4.4 to Figure 4.3, it is seen that the systems  $k_a(t)$  and  $k_b(t)$  have been replaced by simple "through-connections". The response of such through-connection to a unit impulse is the impulse itself, since this connection does not alter a signal at all. From system theory we know that the Fourier transform of a unit impulse is equal to one. Hence, from equation (4.15) we find that the Fourier transform of the system of Figure 4.4 is given by

$$H_2(j\omega_1, j\omega_2) = K_2 \tag{4.16}$$

This means that the second-order kernel of a nonlinear circuit without memory (no capacitive or inductive effects) for which the input-output relationship is described by a power series of the form

$$y(t) = K_1 x(t) + K_2(x(t))^2 + K_3(x(t))^3 + \dots$$
 (4.17)

the second-order transform is  $K_2$ .

#### 4.3.2 Basic third-order system

Similarly to the previous section we consider now a general block diagram representation of a third-order Volterra operator. We will consider a block diagram with linear systems and multipliers that multiply two signals at a time, as we did in the second-order system of Figure 4.3. It is clear that a third-order system will require two such multipliers. In this way a third-order system combines three signals to produce a third-order signal. The most general third-order system with only two multipliers is shown in Figure 4.5. As shown, x(t) is the common input to the second-order system  $\mathbf{F}_2$  with the output  $z_a(t)$  and to the linear system with the impulse response  $k_d(t)$  and output  $z_b(t)$ . The second-order system  $\mathbf{F}_2$  is identical to the one shown in Figure 4.3. This system already combines two signals such that the last multiplier (in front of the linear system with impulse response  $k_e(t)$ ) effectively combines three signals. Once again we remark that

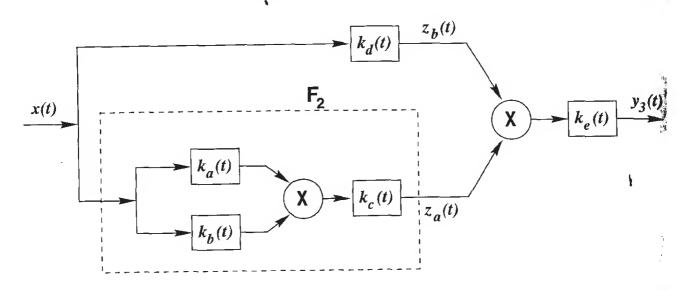


Figure 4.5: Block-diagram representation that illustrates the operation of a simple third-order system.

the parts in Figure 4.5 do not necessarily respond to "physical" parts of a practical third-ord nonlinearity.

In [Sche 80] it is proven that the third-order kernel transform of this system is given by

$$H_3(j\omega_1, j\omega_2, j\omega_3) = K_a(j\omega_1)K_b(j\omega_2)K_c(j\omega_1 + j\omega_2)K_d(j\omega_3)K_e(j\omega_1 + j\omega_2 + j\omega_3)$$
 (4.18)

Let us now simplify the block diagram of Figure 4.5 to a memoryless system. This is done by replacing the blocks corresponding to  $k_a(t)$ ,  $k_b(t)$ ,  $k_c(t)$  and  $k_d(t)$  by through-connection while the block that corresponds to  $k_e(t)$  is replaced by a block that performs a scaling with factor  $K_3$ . The resulting block diagram is shown in Figure 4.6. With the same reasoning as it

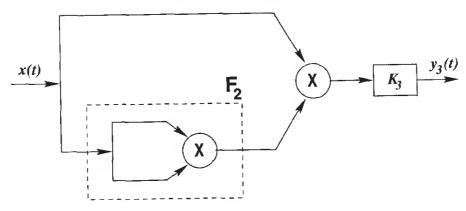


Figure 4.6: Block-diagram representation of a simple memoryless third-order system.

the previous section we find that the third-order kernel transform reduces to

$$H_3(j\omega_1, j\omega_2, j\omega_3) = K_3 \tag{4.1}$$

This result could have been predicted with the knowledge of the previous section.

#### 4.3.3 Application: a nonlinear amplifier

We now analyze the circuit of Figure 4.7: it consists of a nonlinear amplifier with an RC circuit at its input and its output. Capacitive effects inside the amplifier itself are neglected.

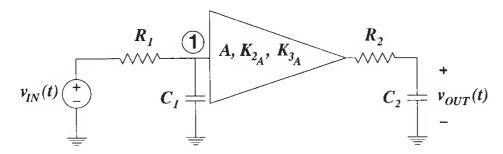


Figure 4.7: A nonlinear amplifier with an RC network at its input and output.

The input impedance of the amplifier is assumed to be infinite and the output impedance is zero. Since capacitive (and inductive) effects in the amplifier are neglected, we can describe the amplifier with nonlinearity coefficients that are independent of frequency. The relationship between the input voltage and the output voltage of the amplifier is given by

$$v_{out}(t) = A v_{in}(t) + K_{2_A} (v_{in}(t))^2 + K_{3_A} (v_{in}(t))^3$$
(4.20)

It is clear that the amplifier nonlinearity is memoryless.

We will handle this example further in the frequency domain. With simple network analysis we find that the ratio between the input voltage  $V_{in}$  and the voltage  $V_1$  at the input of the amplifier is given by

$$\frac{V_1}{V_{in}} = \frac{1}{1 + j\omega R_1 C_1} \tag{4.21}$$

With this result it is easy to see that the transfer function  $H_1(j\omega)$  of the overall linearized circuit is given by

$$H_1(j\omega) = \frac{V_{out}}{V_{in}} = \frac{A}{(1+j\omega R_1 C_1)(1+j\omega R_2 C_2)}$$
 (4.22)

We now consider the second-order kernel transform of the circuit. This kernel is nonzero due to the second-order nonlinearity of the amplifier. This second-order nonlinearity is fed by the voltage over the capacitor  $C_1$  or, in other words, by the output of the RC network  $R_1 - C_1$ . The second-order memoryless nonlinearity of the amplifier can be represented like the block diagram of Figure 4.4. Combining this representation with blocks that represent the RC networks at the input and the output yields the configuration of Figure 4.8 from which the second-order response

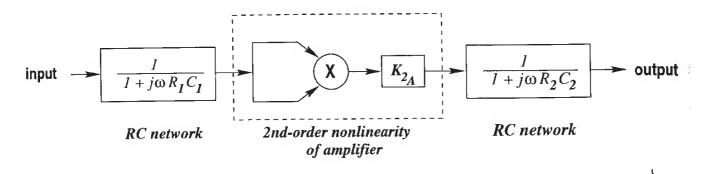


Figure 4.8: Block diagram used for the computation of the second-order response of the circuit of Figure 4.7.

can be computed. Clearly, this block diagram can be transformed to the block diagram of Figure 4.9. The operation of this diagram can be explained easily: the second-order nonlinearity of the amplifier squares the voltage over the capacitor to produce a second-order output signal. This second-order signal then propagates through the RC network at the output. In this block scheme the scale factor  $K_{2_A}$  of the second-order nonlinearity has been combined into one block with the representation of the RC network at the output. This block diagram can be identified with

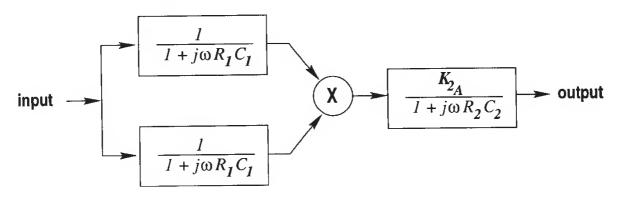


Figure 4.9: Equivalent of the block diagram of Figure 4.8.

the block diagram of Figure 4.3. For the latter we can immediately write down the second-order kernel transform  $H_2(j\omega_1, j\omega_2)$ , using equation (4.15):

$$H_2(j\omega_1, j\omega_2) = \frac{K_{2A}}{(1 + j\omega_1 R_1 C_1) (1 + j\omega_2 R_1 C_1) (1 + j(\omega_1 + \omega_2) R_2 C_2)}$$
(4.23)

Let us now consider the third-order kernel transform which is caused by the third-order non-linearity of the amplifier. This third-order nonlinearity is fed by the voltage over the capacitor  $C_1$ . Using the general block representation of a memoryless third-order nonlinearity that is given in Figure 4.6, the block diagram from which the third-order response can be computed, can be set up easily, as shown in Figure 4.9. This diagram can be transformed into the diagram of Figure 4.11. Identifying this diagram with the general representation of Figure 4.5, we can use

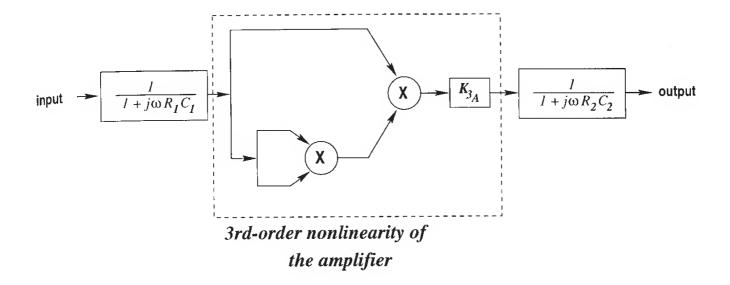


Figure 4.10: Block diagram used for the computation of the third-order response of the circuit of Figure 4.7.

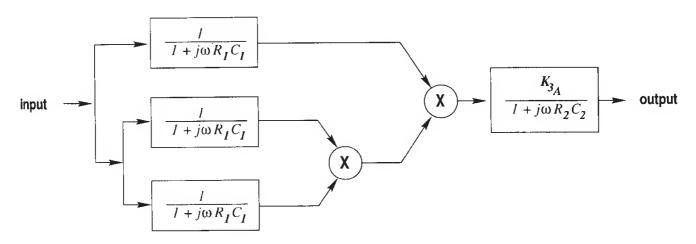


Figure 4.11: Equivalent of the block diagram of Figure 4.10.

equation (4.18) to write down the third-order kernel transform  $H_3(j\omega_1, j\omega_2, j\omega_3)$  of the complete circuit:

$$H_3(j\omega_1, j\omega_2, j\omega_3) = \frac{K_{3_A}}{(1 + j\omega_1 R_1 C_1) (1 + j\omega_2 R_1 C_1) (1 + j\omega_3 R_1 C_1) (1 + j(\omega_1 + \omega_2 + \omega_3) R_2 C_2)}$$
(4.24)

From the second- and third-order kernel transform we can now compute the response to a sinusoidal signal. To this purpose we take a sinusoidal input signal

$$v_{in}(t) = V_{in}\cos(\omega_x t) \tag{4.25}$$

**First-order response** The first-order response  $y_1(t)$  can be found by combining the expression of the linear transfer function, equation (4.22) with the general expression of a linear system to a sinusoidal response, equation (4.13). This yields

$$y_1(t) \approx \frac{V_{in} \cdot A}{\sqrt{1 + \omega_x^2 R_1^2 C_1^2} \sqrt{1 + \omega_x^2 R_2^2 C_2^2}} \cdot \cos(\omega_x t - \tan(\omega_x R_1 C_1) - \tan(\omega_x R_2 C_2))$$
 (4.26)

The  $\approx$  symbol in equation (4.26) is due to the fact that the response at the frequency  $\omega_x$  is not only caused by the linear behavior as indicated in this equation, but also by third-order and higher-order behavior as we saw before.

Second-order response For the computation of the second-order response  $y_2(t)$  to the sinusoidal signal of equation (4.25) we combine the general equation for the response of a second-order system to a sinusoidal signal (equation (4.11)) with the second-order kernel transform of the circuit (see equation (4.23)). This yields after some algebra

$$y_{2}(t) = \frac{V_{in}^{2} \cdot K_{2_{A}}}{2\left(1 + \omega_{x}^{2} R_{1}^{2} C_{1}^{2}\right) \sqrt{1 + 4\omega_{x}^{2} R_{2}^{2} C_{2}^{2}}} \cdot \left[1 + \cos\left(2\omega_{x} t - 2\operatorname{atan}(\omega_{x} R_{1} C_{1}) - \operatorname{atan}(2\omega_{x} R_{2} C_{2}\right)\right] + \frac{V_{in}^{2} K_{2_{A}}}{2\left(1 + \omega_{x}^{2} R_{1}^{2} C_{1}^{2}\right)}$$

$$(4.27)$$

Clearly, this response is seen to consist of a DC shift and a second harmonic. The magnitude of the two components depends on the frequency  $\omega_x$ . Note that the DC shift is independent of  $R_2C_2$ . The reason is that this DC shift is a signal that is produced by the second-order nonlinearity of the amplifier. This signal propagates at 0Hz through the RC network at the output without attenuation.

The second harmonic distortion can now be found by taking the ratio of the response at  $2\omega_x$  and at  $\omega_x$ . This yields

$$HD_2 = \frac{V_{in}}{2} \cdot \frac{K_{2_A}}{A} \cdot \left| \frac{1}{1 + j\omega_x R_1 C_1} \cdot \frac{1 + j\omega_x R_2 C_2}{1 + j2\omega_x R_2 C_2} \right|$$
(4.28)

$$= \frac{V_{in}}{2} \cdot \frac{K_{2_A}}{A} \cdot \frac{1}{\sqrt{1 + \omega_x^2 R_1^2 C_1^2}} \cdot \sqrt{\frac{1 + \omega_x^2 R_2^2 C_2^2}{1 + 4\omega_x^2 R_2^2 C_2^2}}$$
(4.29)

It is seen that the second harmonic distortion depends on the normalized second-order nonlinearity coefficient  $K_{2_A}/A$  of the amplifier. Further it is seen that the second harmonic distortion decreases with increasing frequency. This is due to the fact that the second harmonic decrease faster with increasing frequency than the fundamental response, since the former contains the square of  $(1+j\omega_xR_1C_1)$  in the denominator, while the latter contains only the first power of  $(1+j\omega_xR_1C_1)$  in its denominator. In addition, the second harmonic is a component of a second order signal that is produced by the second-order nonlinearity and propagates at a frequency  $2\omega$  to the output, whereas the first-order response is due to the input signal that propagates at a frequency  $\omega_x$  in a linear way to the output. If  $\omega_x \gg 1/(R_2C_2)$  then the ratio of the second harmonic and the first-order response is a factor of 6dB smaller than at low frequencies.

**Third-order response** The third-order response  $y_3(t)$  can be found by combining the general equation for a third-order system to a sinusoidal signal, equation (4.12), with the expression of the third-order kernel transform of the circuit, equation (4.24). This yields a response at  $3\omega_x$  and at  $\omega_x$ :

$$y_{3}(t) = \frac{V_{in}^{3} \cdot K_{3_{A}}}{4\left(1 + \omega_{x}^{2}R_{1}^{2}C_{1}^{2}\right)^{3/2}\sqrt{1 + 9\omega_{x}^{2}R_{2}^{2}C_{2}^{2}}}\cos\left(3\omega_{x}t - 3\tan(\omega_{x}R_{1}C_{1}) - \tan(3\omega_{x}R_{2}C_{2})\right) + \frac{3V_{in}^{3}K_{3_{A}}}{4\left(1 + \omega_{x}^{2}R_{1}^{2}C_{1}^{2}\right)^{3/2}\sqrt{1 + \omega_{x}^{2}R_{2}^{2}C_{2}^{2}}}\cos\left(\omega_{x}t - \tan(\omega_{x}R_{1}C_{1}) - \tan(\omega_{x}R_{2}C_{2})\right)$$

$$(4.30)$$

The third harmonic distortion is then found to be

$$HD_{3} = \frac{V_{in^{2}}}{4} \cdot \frac{K_{3_{A}}}{A} \cdot \left| \frac{1}{(1+j\omega_{x}R_{1}C_{1})^{2}} \cdot \frac{1+j\omega_{x}R_{2}C_{2}}{1+j3\omega_{x}R_{2}C_{2}} \right|$$
(4.31)

$$= \frac{V_{in^2}}{4} \cdot \frac{K_{3_A}}{A} \cdot \frac{1}{1 + \omega_x^2 R_1^2 C_1^2} \cdot \sqrt{\frac{1 + \omega_x^2 R_2^2 C_2^2}{1 + 9\omega_x^2 R_2^2 C_2^2}}$$
(4.32)

The interpretation of  $HD_3$  is similar to the interpretation of  $HD_2$ .

Intermodulation products can also be computed with the knowledge of the kernel transforms of different order. This is postponed until the next section, after we defined the performance parameters discussed in Chapter 2, in terms of the kernel transforms.

# 4.4 Nonlinear performance parameters in terms of Volterra kernels

The analytic description of a nonlinear system with Volterra series allows to extend the definitions from Chapter 2 with frequency dependencies.

#### 4.4.1 Single-tone and two-tone definitions

When a system that can be described by a Volterra series up to order three, is excited by the sum of two sinusoidal excitations  $A_1 \cos \omega_1 t$  and  $A_2 \cos \omega_2 t$ , then the output is given by the sum of the responses listed in Table 4.1. In Appendix B it is shown how these responses can be computed in terms of Volterra kernels. The total output consists of 18 responses, at 13 different frequencies. In Table 4.1 these responses are classified as linear responses, harmonics, intermodulation products, desensitizations and compressions or expansions, according to the definitions given in Chapter 2.

It is interesting to compare the responses listed in Table 4.1 to the responses of a memoryless circuit, as given in Figure 2.6. We already found in Section 4.3 that the second- and third-order kernel transform of a memoryless system are equal to the second- and third-order coefficients  $K_2$  and  $K_3$  of the power series expansion of the input-output relationship of the circuit. These coefficients  $K_2$  and  $K_3$  have been used in Figure 2.6 to express the responses. In fact, the responses

order	frequency of response	amplitude of response	type of response
1	$\omega_1$	$A_1  H_1(j\omega_1) $	linear
1	$\omega_2$	$A_2 \left  H_1(j\omega_2) \right $	1
2	$\omega_1 + \omega_2$	$A_1A_2  H_2(j\omega_1,j\omega_2) $	2nd-order intermodulation products
2	$ \omega_1 - \omega_2 $	$A_1 A_2  H_2(j\omega_1, -j\omega_2) $	
2	$2\omega_1$	$\frac{1}{2}A_1^2\left H_2(j\omega_1,j\omega_1)\right $	2nd harmonics
2	$2\omega_2$	$\frac{1}{2}A_2^2 \left  H_2(j\omega_2, j\omega_2) \right $	
2	0	$rac{1}{2}A_1^2\left H_2(j\omega_1,-j\omega_1) ight $	DC shift
2	0	$\frac{1}{2}A_2^2 \left  H_2(j\omega_2, -j\omega_2) \right $	DC simt
3	$2\omega_1 + \omega_2$	$\frac{3}{4}A_1^2A_2  H_3(j\omega_1, j\omega_1, j\omega_2) $	third-order
3	$ 2\omega_1-\omega_2 $	$\left  \frac{3}{4}A_{1}^{2}A_{2}\left  H_{3}(j\omega_{1},j\omega_{1},-j\omega_{2}) \right  \right $	intermodulation
3	$\omega_1 + 2\omega_2$	$\frac{3}{4}A_1A_2^2  H_3(j\omega_1, j\omega_2, j\omega_2) $	products
3	$ \omega_1-2\omega_2 $	$\frac{3}{4}A_1A_2^2 H_3(j\omega_1, -j\omega_2, -j\omega_2) $	p. 0 3 3 3 3
3	$\omega_1 + \omega_2 - \omega_2 = \omega_1$	$\frac{3}{2}A_1A_2^2 H_3(j\omega_1,j\omega_2,-j\omega_2) $	third-order
3	$\omega_1 - \omega_1 + \omega_2 = \omega_2$	$\frac{3}{2}A_1^2A_2 H_3(j\omega_1,-j\omega_1,j\omega_2) $	desensitization
3	$2\omega_1 - \omega_1 = \omega_1$	$rac{3}{4}A_1^3\left H_3(j\omega_1,j\omega_1,-j\omega_1) ight $	third-order compression
3	$2\omega_2 - \omega_2 = \omega_2$	$\left  \begin{array}{c} rac{3}{4}A_2^3 \left  H_3(j\omega_2, j\omega_2, -j\omega_2)  ight  \end{array} \right $	or expansion
3	$3\omega_1$	$\left  \frac{1}{4}A_1^3 \left  H_3(j\omega_1, j\omega_1, j\omega_1) \right  \right $	third harmonics
3	$3\omega_2$	$\frac{1}{4}A_2^3 \left  H_3(j\omega_2, j\omega_2, j\omega_2) \right $	

Table 4.1: Different responses at the output of a nonlinear system described by Volterra kernels of order one, two and three, excited by two sinusoids  $A_1 \cos \omega_1 t$  and  $A_2 \cos \omega_2 t$ . The integer in the first column indicates the order of the kernel by which the response is determined.

given in Figure 2.6 can be obtained from Table 4.1 by evaluating the kernel transforms at very low frequencies, i.e. for both  $j\omega_1$  and  $j\omega_2$  equal to zero. For example, the third-order desensitization and the third-order compression or expansion are both proportional to  $H_3(0,0,0)$  and when the input amplitudes  $A_1$  and  $A_2$  are both equal to A, then they combine to  $\frac{9}{4}A^3H_3(0,0,0)=\frac{9}{4}A^3K_3$ 

which is exactly what we found in Figure 2.6.

**Harmonic distortion** In Section 4.3.3 we already computed the harmonic distortion figures for a specific example. The expressions for the second and third harmonic distortion in terms of general Volterra are easily found either from Table 4.1 or from equations (4.11) and (4.12):

$$HD_2 = \frac{A_1}{2} \left| \frac{H_2(j\omega_1, j\omega_1)}{H_1(j\omega_1)} \right| \tag{4.33}$$

$$HD_3 = \frac{A_1^2}{4} \left| \frac{H_3(j\omega_1, j\omega_1, j\omega_1)}{H_1(j\omega_1)} \right|$$
(4.34)

Note that we already used these definitions in Section 4.3.3.

Intermodulation distortion The intermodulation distortion figures  $IM_2$  and  $IM_3$  can be computed by taking ratios of the appropriate signals from Table 4.1. Assume for example that the wanted output signal of a circuit is the signal at frequency  $\omega_1$  and that we want to specify the intermodulation distortion for the intermodulation product at the sum frequency  $\omega_1 + \omega_2$ . Then the intermodulation distortion  $IM_2$  is found by taking the amplitude of the input signals equal and referring the intermodulation product to the output at  $\omega_1$  (see also equation (2.26)). This yields

$$IM_2 = A_1 \left| \frac{H_2(j\omega_1, j\omega_2)}{H_1(j\omega_1)} \right|$$
 (4.35)

and for the intermodulation product at the difference frequency

$$IM_2 = A_1 \left| \frac{H_2(j\omega_1, -j\omega_2)}{H_1(j\omega_1)} \right|$$
 (4.36)

It is seen that now two definitions for  $IM_2$  exist, depending on which second-order intermodulation product is considered. For memoryless systems one can do with only one definition that applies to both intermodulation products. The reason is that for memoryless systems  $H_2(j\omega_1, j\omega_2) = H_2(0,0) = H_2(j\omega_1, -j\omega_2)$ .

Third-order intermodulation distortion figures are defined in a similar way. In RF applications where the useful signal is at a frequency  $\omega_1$  and an unwanted signal is present at a frequency  $\omega_2$  close to  $\omega_1$  the intermodulation product at  $2\omega_2 - \omega_1$  can give rise to crosstalk. Hence it is interesting to relate this unwanted intermodulation product with the wanted signal. Using Table 4.1 the third-order intermodulation distortion in terms of Volterra kernel transforms

$$IM_3 = \frac{3}{4} A_1^2 \left| \frac{H_3(j\omega_1, -j\omega_2, -j\omega_2)}{H_1(j\omega_1)} \right|$$
 (4.37)

or, using equation (4.10)

$$IM_3 = \frac{3}{4} A_1^2 \left| \frac{H_3(-j\omega_1, j\omega_2, j\omega_2)}{H_1(j\omega_1)} \right|$$
(4.38)

It should be noted that for third-order intermodulation products at different frequencies another definition applies, which differs from equation (4.38) in the arguments of  $H_3$ .

In Chapter 2 we found that for weakly nonlinear systems without memory the ratio  $IM_2/HD_2$  equals two and  $IM_3/HD_3$  equals three (see equations (2.31) and (2.32)). Using equations (4.33), (4.34), (4.35) and (4.38) these ratios in terms of Volterra kernel transforms are given by

$$\frac{IM_2}{HD_2} = 2 \left| \frac{H_2(j\omega_1, j\omega_2)}{H_2(j\omega_1, j\omega_1)} \right| \tag{4.39}$$

for the intermodulation product at  $\omega_1 + \omega_2$  and

$$\frac{IM_3}{HD_3} = 3 \left| \frac{H_3(-j\omega_1, j\omega_2, j\omega_2)}{H_3(j\omega_1, j\omega_1, j\omega_1)} \right|$$
(4.40)

for the third-order intermodulation product at  $|2\omega_2 - \omega_1|$ . For memoryless systems  $H_2$  and  $H_3$  are the same for any frequency argument such that again  $IM_2/HD_2$  and  $IM_3/HD_3$  equal 2 and 3, respectively. Equivalently, this also applies at very low frequencies, where  $\omega_1, \omega_2 \to 0$ . However, for systems with memory this no longer holds. This can be illustrated with the circuit with the nonlinear amplifier from Section 4.3.3.

For this circuit the ratio  $IM_3/HD_3$  is found by combining equation (4.24) with equation (4.40). This yields

$$\frac{IM_3}{HD_3} = \left| \frac{(1+j\omega_1 R_1 C_1)^3 (1+j3\omega_1 R_2 C_2)}{(1-j\omega_1 R_1 C_1) (1+j\omega_2 R_1 C_1)^2 (1+j(2\omega_2 - \omega_1) R_2 C_2)} \right|$$
(4.41)

Assume that the time constant  $R_1C_1$  is much smaller than the product  $R_2C_2$ . Assume further that  $\omega_1 \approx \omega_2 = \omega_x \gg 1/(R_2C_2)$  and  $\omega_x \ll 1/(R_1C_1)$ . Then  $IM_3/HD_3$  is approximately equal to

$$\frac{IM_3}{HD_3} \approx 9 \tag{4.42}$$

It is seen that  $IM_3$  is larger than  $3HD_3$  in this case. This is explained by the fact that for a single-tone excitation the third-order nonlinearity produces a third-order signal that propagates through the rest of the circuit at frequency  $3\omega_1$ . On the other hand, the third-order intermodulation product at  $2\omega_2-\omega_1$  produced by the third-order nonlinearity propagates through the rest of the circuit at a lower frequency since here  $2\omega_2-\omega_1<3\omega_1$ . If the part of the circuit after the third-order nonlinearity has a low-pass characteristic, then the transfer of a signal from the nonlinearity to the output at  $3\omega_1$  is smaller than at  $2\omega_2-\omega_1$ .

**Intercept points** The intercept points  $IP_{2h}$  and  $IP_{3h}$  are the amplitudes for which  $HD_2$  and  $HD_3$  respectively become one:

$$IP_{2h} = 2 \left| \frac{H_1(j\omega_1)}{H_2(j\omega_1, j\omega_1)} \right|$$
 (4.43)

$$IP_{3h} = 2\sqrt{\left|\frac{H_1(j\omega_1)}{H_3(j\omega_1, j\omega_1, j\omega_1)}\right|}$$
 (4.44)

Intercept points for intermodulation products are defined similarly. Of course, the different definitions in terms of Volterra kernels that are needed depending on the frequency of the intermodulation product, give rise to different definitions for intercept points. For example, for the intermodulation product at  $2\omega_2 - \omega_1$  we find  $IP_{3i}$  as the amplitude for which  $IM_3 = 1$ . From equation (4.38) we find then

$$IP_{3i} = \frac{2}{\sqrt{3}} \sqrt{\frac{H_1(j\omega_1)}{H_3(-j\omega_1, j\omega_2, j\omega_2)}}$$
(4.45)

#### 4.4.2 **Cross modulation**

In Chapter 2 it was shown how at low frequencies amplitude modulation of a carrier can be transferred to another carrier. At high frequencies, however, phase modulation can occur as well. In [Mey 72] it is shown that when the input x(t) to a weakly nonlinear circuit described by a Volterra series is an amplitude-modulated signal at  $\omega_1$  together with an unmodulated carrier with frequency  $\omega_2$ :

$$x(t) = A(1 + m_1 \cos \omega_m t) \cos \omega_1 t + A \cos \omega_2 t \tag{4.46}$$

in which  $\omega_m \ll \omega_1, \omega_2$ , then the output terms at frequencies  $\omega_2$  and  $\omega_2 \pm \omega_m$  are given by

$$y(t) = A |H_1(j\omega_2)| \left[\cos(\omega_2 t + \beta_1) + m_x \cos(\omega_2 t + \beta_2) \cos\omega_m t\right]$$
(4.47)

in which

$$m_x = 3m_1 \frac{|H_3(j\omega_1, -j\omega_1, j\omega_2)|}{|H_1(j\omega_2)|} A^2$$
(4.48)

and  $\beta_1$  and  $\beta_2$  are the phase angles of  $H_1(j\omega_2)$  and  $H_3(j\omega_1, -j\omega_1, j\omega_2)$ , respectively. Setting  $\omega_1 = \omega_2 = 0$  in equation (4.47) leads to the same conclusions as in Section 2.6.

Let  $\beta_2 - \beta_1 = \phi$ , then for  $m_x \ll 1$  the output terms in equation (4.48) can be expressed as

$$y(t) = A |H_1(j\omega_2)| (1 + m_x \cos\phi \cos\omega_m t) \cos(\omega_2 t + \beta_1 + m_x \sin\phi \cos\omega_m t)$$
 (4.49)

Hence, both phase and amplitude cross modulation can occur, depending on  $\phi$ . If  $\phi = 0$  then no phase cross modulation occurs, and when  $\phi = \pi/2$ , then there is pure phase cross modulation. Whereas in Section 2.6 only one cross-modulation factor was defined, since only amplitude cross modulation occurs at low frequencies, it is appropriate to define two factors for high-frequency cross modulation. The amplitude- and phase cross-modulation factor  $CM_A$  and  $CM_P$  are defined as the transferred amplitude modulation and phase modulation, respectively, divided by the original modulation index:

$$CM_A = \frac{m_x \cos \phi}{m_1} \tag{4.50}$$

$$CM_A = \frac{m_x \cos \phi}{m_1}$$

$$CM_P = \frac{m_x \sin \phi}{m_1}$$
(4.50)

At low frequencies we found in equation (2.44) that the ratio of  $CM_A$  and third-order intermodiation distortion is four. This is no longer true at high frequencies.

Power series do not take into account phase information such that AM to PM conversion cannot be explained. In this section it has been shown that the Volterra series approach does take into account phase information.

#### 4.5 Suppression of even-order or odd-order kernels

In Section 2.3 we mentioned that the output of a balanced circuit does not contain even-order harmonics or intermodulation products when the input ports of this circuit are excited by two input signals of equal amplitude and opposite phase. In this section, this result is generalized with the use of Volterra series.

Figure 4.12 and 4.13 each show a nonlinear system.

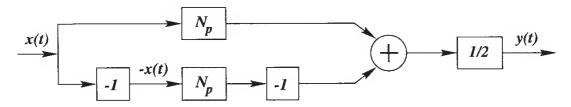


Figure 4.12: Connection of two identical pth-order systems to suppress the response resulting from the even-order kernels of the original system  $N_p$ .

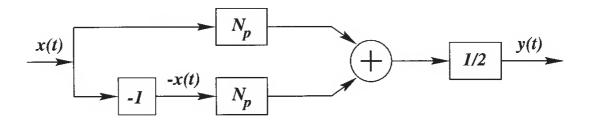


Figure 4.13: Connection of two identical pth-order systems to suppress the response resulting from the odd-order kernels of the original system  $N_p$ .

The response of the system in Figure 4.12 does not contain even-order signals. Conversely the response of the system in Figure 4.13 does not contain odd-order signals. Also, the fundamental response is suppressed. This can be useful in circuits such as balanced or double-balanced mixers.

The construction of these two systems is based upon the following theorem.

**THEOREM 4.1.** If for all input signals x(t) for which the Volterra series of a nonlinear system converges, the output y(x(t)) satisfies the relationship

$$y(x(t)) = -y(-x(t))$$

then  $h_{2n}(\tau_1, \ldots, \tau_{2n}) = 0$  for all  $\tau_1, \ldots, \tau_{2n}$  and  $n = 1, 2, \ldots$ If for all input signals x(t) for which the Volterra series of a nonlinear system converges, the output y(x(t)) satisfies the relationship

$$y(x(t)) = y(-x(t))$$

then 
$$h_{2n+1}(\tau_1,\ldots,\tau_{2n+1})=0$$
 for all  $\tau_1,\ldots,\tau_{2n+1}$  and  $n=0,1,2,\ldots$ 

This theorem reformulates the Theorems 2.1 and 2.2 in terms of Volterra series. The suppression of the response of even-order (odd-order) kernels allows to suppress all even-order (odd-order) harmonics or intermodulation products at the output of a circuit.

The first part of Theorem 4.1 is applicable to the block diagram of Figure 4.12 whereas the second part applies for the system of Figure 4.13. Indeed, the output of the system of Figure 4.12 as a result of an input signal x(t) is given by

$$y_a(t) = \frac{1}{2} \left( -\mathbf{N}_p[-x(t)] + \mathbf{N}_p[x(t)] \right)$$
 (4.52)

whereas the response to -x(t) is given by

$$y_b(t) = \frac{1}{2} \left( \mathbf{N}_p[-x(t)] - \mathbf{N}_p[x(t)] \right)$$
 (4.53)

which is seen to be the opposite of  $y_a(t)$ .

Similarly, the response of the system of Figure 4.13 to both x(t) and -x(t) is given by

$$y_c(t) = \frac{1}{2} \left( \mathbf{N}_p[-x(t)] + \mathbf{N}_p[x(t)] \right)$$
 (4.54)

The block diagram of Figure 4.12 is the principle schematic of a differential circuit: two opposite signals x(t) and -x(t) are applied to two identical blocks, represented by  $N_p$ . The difference between the output of these two blocks is the overall output. Hence, Theorem 4.1 gives a theoretical explanation of the success of differential circuits for the suppression of even-order nonlinear responses.

The suppression of odd-order kernels can be of practical interest as well, for example in multipliers and mixers. In the schematic of Figure 4.13 the linear response, which is the first-order response is suppressed as well. This principle is used in a double-balanced mixer. For example, when a mixer is used as a downconverter, then the mixer inputs are the RF signal and the local oscillator signal. It is often unwanted that these signals are seen at the mixer output. The propagation of these signals through the mixer are mainly due to first-order behavior, but also to higher odd-order behavior. In order to suppress these signals at the output, the principle schematic of Figure 4.13 is used for the mixer.

#### 4.5.1 Application: suppression in a differential pair

A basic building block of analog integrated circuits is a differential pair. A bipolar differential pair is shown in Figure 4.14. The input signal is a differential voltage source. It is assumed that the applied input signal is small enough such that no Volterra series that describes a voltage or

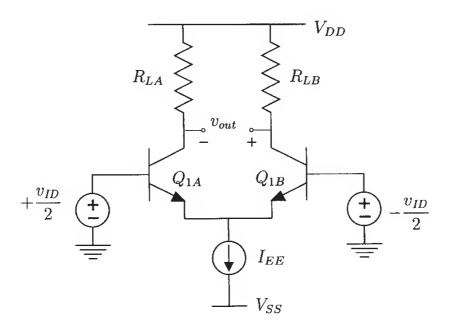


Figure 4.14: A simple differential pair.

current in the circuit, diverges. The output of the circuit is the difference of the collector voltages. It is assumed that the transistors and the resistors match perfectly.

The differential pair does not literally correspond to the principle schematic of Figure 4.12, in the sense that every transistor loaded with a resistor could be associated with a block  $N_p$  from Figure 4.12. Indeed, the input signal with amplitude  $+v_{ID}/2$  that is applied at the base of  $Q_{1A}$  does not only propagate through transistor  $Q_{1A}$ , but also along the common emitter through  $Q_{1B}$  to the output. A similar reasoning holds for the input signal  $-v_{ID}/2$ . If the common emitter would be a real ground, then the circuit will not be able to distinguish common-mode signals from differential ones, such that a big advantage of the differential mode of operation is lost. Nevertheless, a similarity with the principle of Figure 4.12 can be seen: the voltage source with value  $+v_{ID}/2$  sees the same nonlinear circuitry (corresponding to the operator  $N_p$  in Figure 4.12) as the source with value  $-v_{ID}/2$ .

The suppression of even-order responses can also be investigated using the explicit expression for the input-output relationship of the circuit. We will derive this relationship without taking into account frequency effects. Such relationship is often referred to as a DC transfer characteristic. In the derivation of this characteristic we will further neglect all ohmic resistances and output conductances of the transistor, as well as the output conductance of the current source at the common emitter.

Summing the voltages around the loop consisting of the two voltage sources and the two base-emitter voltages yields

$$\frac{v_{ID}}{2} - v_{BE_{1A}} + v_{BE_{1B}} + \frac{v_{ID}}{2} = 0 {(4.55)}$$

The base-emitter voltages can be expressed as a function of the collector current through  $Q_{1A}$ 

and  $Q_{1B}$  using the simplified model of the collector current (equation (3.12))

$$v_{BE_{1A}} = V_t \ln \frac{i_{C_{1A}}}{I_{S_{1A}}} \tag{4.56}$$

$$v_{BE_{1B}} = V_t \ln \frac{i_{C_{1B}}}{I_{S_{1B}}} \tag{4.57}$$

Putting  $I_{S_{1A}} = I_{S_{1B}} = I_S$  we find by combining equations (4.55), (4.56) and (4.57)

$$\frac{i_{C_{1A}}}{i_{C_{1AB}}} = \exp\left(\frac{v_{ID}}{V_t}\right) \tag{4.58}$$

The current  $I_{EE}$  at the common emitter point is the sum of the emitter currents. This means that

$$I_{EE} = \frac{\beta}{\beta + 1} \left( i_{C_{1A}} + i_{C_{1B}} \right) \tag{4.59}$$

Combining equations (4.58) and (4.59) then yields

$$i_{C_{1A}} = \frac{\beta I_{EE}}{(\beta+1)\left(1+\exp\left(-\frac{v_{ID}}{V_t}\right)\right)} \tag{4.60}$$

$$i_{C_{1B}} = \frac{\beta I_{EE}}{(\beta + 1)\left(1 + \exp\left(\frac{v_{ID}}{V_t}\right)\right)}$$
(4.61)

The output voltage  $v_{OUT}$  is given by

$$v_{OUT} = R_{LB}i_{C_{1B}} - R_{LA}i_{C_{1A}} (4.62)$$

Using equations (4.60) and (4.61) and putting  $R_{LA} = R_{LB} = R_L$  we finally obtain

$$v_{OUT} = I_{EE} R_L \frac{\beta}{\beta + 1} \tanh\left(\frac{v_{ID}}{V_t}\right) \tag{4.63}$$

This output voltage is shown in Figure 4.15 as a function of the input voltage. It is seen that the output voltage is an odd function of the input voltage  $v_{ID}$ . For such function, the conditions of Theorem 2.1 apply. Hence, no even harmonics occur in the output voltage.

Consider now the voltage  $v_E$  at the common emitter of the differential pair of Figure 4.14. This is found to be

$$v_E = \frac{v_{ID}}{2} - V_t \ln \frac{\beta}{\beta + 1} \frac{I_{EE}}{I_S} + V_t \ln \left( 1 + \exp\left(-\frac{v_{ID}}{V_t}\right) \right)$$
(4.64)

This voltage is depicted in Figure 4.16. It is seen that the voltage at the common-emitter point is an even function of the input voltage since this voltage moves in exactly the same way if the roles of  $+v_{ID}/2$  and  $-v_{ID}/2$  are interchanged. Hence, according to Theorem 2.2, no odd-order harmonics occur at this point. This corresponds with the assumption that the common-emitter point is an AC ground when the circuit is linearized: the presence of this AC ground means that no fundamental response is present at the common emitter point.

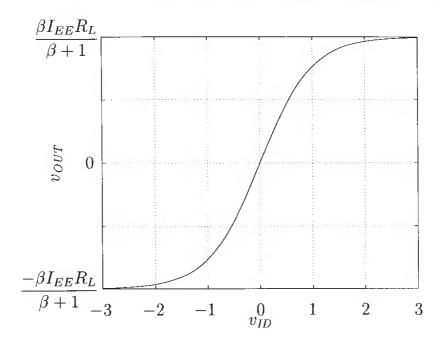


Figure 4.15: Differential output voltage of the differential pair from Figure 4.14 as a function of the differential input voltage.

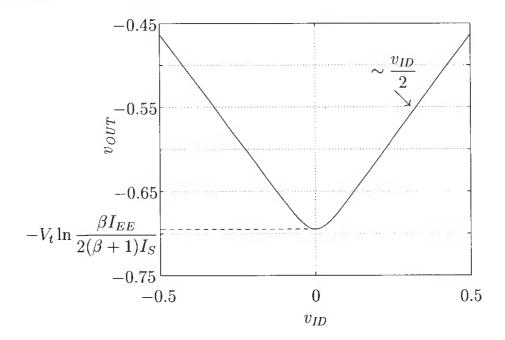


Figure 4.16: Voltage at the common emitter point of the differential pair from Figure 4.14 as a function of the differential input voltage.

#### 4.5.2 Application: suppression in a Gilbert multiplier

In Figure 4.17 a Gilbert multiplier [Gilb 68, Gilb 74] is shown. The inputs to this circuit are the differential voltages  $v_{IN1}$  and  $v_{IN2}$ . It is assumed that these signals are small enough such that the circuits behaves in a weakly nonlinear way. When this circuit is used as an upconverter or a downconverter, then the response of interest when the inputs are sinusoidal, is at the sum

or difference frequency. Ideally, this response should be caused by second-order second-order behavior only. In reality, higher even-order nonlinear behavior also contributes to the responses at the sum and difference frequency. These higher-order contributions cannot be suppressed by using the principle schematic of Figure 4.12 since otherwise the wanted second-order response would be suppressed as well.

On the other hand, the output should not contain a component at the same frequency of one of the two input signals. These frequency components can be suppressed by combining a fully differential operation with the principle from Figure 4.13. Indeed, consider for example the operation of transistors  $Q_{1A}$  and  $Q_{1C}$ . Their inputs, both at the base (due to  $v_{IN1}$ ) and at the emitter (due to  $v_{IN2}$ ) are opposite. The outputs of the transistors, in this case their collector current, are summed, just as in Figure 4.13.

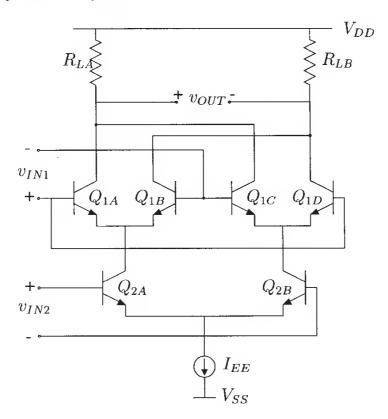


Figure 4.17: A Gilbert multiplier.

## 4.6 Cascade connection of nonlinear systems

In this section we will calculate the Volterra kernels of a system that consists of the cascade connection of two nonlinear blocks. An example of such system is a two-stage amplifier.

In Figure 4.18 the nonlinear system  $\mathbf{H}$  is connected in cascade with the nonlinear system  $\mathbf{F}$ , resulting in the overall system  $\mathbf{Q}$ . We will now compute the Volterra kernels of order one to three of the system  $\mathbf{Q}$ . For the computations it is assumed that the blocks  $\mathbf{H}$  and  $\mathbf{H}$  do not load each

other. While this assumption is often not valid, the idealization yields some interesting insights as pointed out below.

$$\begin{array}{c|c} x(t) & \hline \\ H & \hline \\ \end{array} \begin{array}{c|c} y(t) & \hline \\ F & \hline \end{array} \begin{array}{c|c} z(t) = \mathbf{Q}[x(t)] \\ \hline \end{array}$$

Figure 4.18: The system **Q** formed by the cascade connection of **H** and **F**.

#### 4.6.1 General expressions

In [Sche 80] expressions for the Volterra operators of different order of Q in terms of Volterra operators of the systems H and F are derived. The Volterra operators of order one up to order three are given by

$$\mathbf{Q}_1 = \mathbf{F}_1 \mathbf{H}_1 \tag{4.65}$$

$$\mathbf{Q}_2 = \mathbf{F}_1 \mathbf{H}_2 + \mathbf{F}_2 \mathbf{H}_1 \tag{4.66}$$

$$\mathbf{Q}_3 = \mathbf{F}_1 \mathbf{H}_3 + \mathbf{F}_2 [\mathbf{H}_1 + \mathbf{H}_2] - \mathbf{F}_2 \mathbf{H}_1 - \mathbf{F}_2 \mathbf{H}_2 + \mathbf{F}_3 \mathbf{H}_1$$
 (4.67)

In these expressions, an operator notation of the form  $\mathbf{F}_m \mathbf{H}_n$  means that operator  $\mathbf{F}_m$  acts on the output of  $\mathbf{H}_n$ . Also note that in the expression for  $\mathbf{Q}_3$ , the notation  $\mathbf{F}_2[\mathbf{H}_1 + \mathbf{H}_2]$  is not the same as  $\mathbf{F}_2\mathbf{H}_1 + \mathbf{F}_2\mathbf{H}_2$ .

An easier interpretation can be obtained from the Laplace transforms of these expressions, which are found to be

$$Q_{1}(s_{1}) = F_{1}(s_{1})H_{1}(s_{1})$$

$$Q_{2}(s_{1}, s_{2}) = F_{1}(s_{1} + s_{2})H_{2}(s_{1}, s_{2}) + F_{2}(s_{1}, s_{2})H_{1}(s_{1})H_{1}(s_{2})$$

$$Q_{3}(s_{1}, s_{2}, s_{3}) = F_{3}(s_{1}, s_{2}, s_{3})H_{1}(s_{1})H_{1}(s_{2})H_{1}(s_{3}) + F_{1}(s_{1} + s_{2} + s_{3})H_{3}(s_{1}, s_{2}, s_{3})$$

$$+ \frac{2}{3} \left[ F_{2}(s_{1}, s_{2} + s_{3})H_{1}(s_{1})H_{2}(s_{2}, s_{3}) + F_{2}(s_{2}, s_{1} + s_{3})H_{1}(s_{2})H_{2}(s_{1}, s_{3}) + F_{2}(s_{3}, s_{1} + s_{2})H_{1}(s_{3})H_{2}(s_{1}, s_{2}) \right]$$

$$+ F_{2}(s_{3}, s_{1} + s_{2})H_{1}(s_{3})H_{2}(s_{1}, s_{2})$$

$$(4.70)$$

In the expression of the first-order transform  $Q_1(s_1)$  it is seen that the role of both first-order transfer functions is identical. This means that for linear systems the overall input-output relationship is unaffected by the order of connecting the two systems, at least if there are no loading effects<sup>1</sup>.

For higher-order systems the order of connection does play a role. This is evidenced by the expressions of the second- and third-order kernel transform of **Q**, where the nonlinear transfer

<sup>&</sup>lt;sup>1</sup>This is no longer true if the systems have multiple inputs and outputs.

functions that correspond to H play a different role than the transfer functions corresponding to F.

Consider now the expression of the second-order kernel transform. This consists of two contributions: the first one comes from the second-order nonlinearity of **H** that produces a second-order signal. This signal is fed into **F**. Since we are looking at the second-order transform we do not consider the influence of the nonlinearities of **F** since these nonlinearities produce with this second-order signal a signal that is at least one order higher. Hence we only need to consider the "linear propagation" through **F**. This is clarified in Figure 4.19.

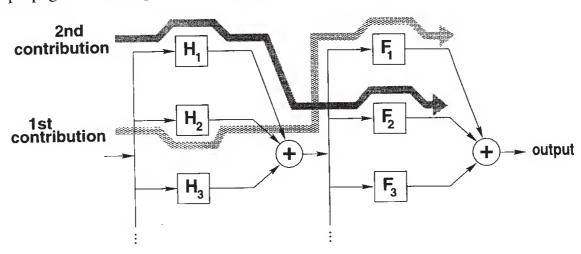


Figure 4.19: Representation of the cascade connection of Figure 4.18 with Volterra operators of different order. The arrows indicate the two contributions to the second-order kernel of the overall system.

The second contribution to the second-order kernel transform is also shown in Figure 4.19. It is caused by the second-order nonlinearity of the second stage that combines two first-order signals that come from the first block, to produce a second-order signal. In the case of an excitation with a single sinusoidal signal, the two first-order signals that come out of **H** are identical. When two different sinusoidal signals are applied then, in addition, combinations of two different signals need to be considered as well.

We now consider the expression for the third-order kernel transform of  $\mathbf{Q}$ . The first two terms of equation (4.70) are similar to the terms of the second-order output. Indeed, the first term reveals that the second stage third-order kernel acts on the fundamental frequencies linearly amplified at the respective frequencies by the first stage. The second term indicates that the third-order output of the first stage at the frequency  $s_1 + s_2 + s_3$  is linearly amplified by the last stage.

The last term of equation (4.70) between the square brackets is explained as follows: the second-order kernel of the last stage acts on the sum of the input signal that is linearly amplified by the first stage and the second-order output of the first stage. In other words, the second-order nonlinearity of the second stage combines two signals and the order of the resulting signal is equal to the sum of the order of the two contributing signals. The term consists of three contributions. While these originate from the same effect, their significance is to make the function

 $Q_3(s_1, s_2, s_3)$  symmetric in each of its arguments. In some works [VdEi 89, Kuo 77] only one term is given for the interaction of the two second-order nonlinearities. Then, however, the third-order transfer function is not symmetric.

Whereas the symmetrization explains the factor 3 in the denominator of the common  $\frac{2}{3}$  factor, the factor 2 is explained by the fact that there are two ways to combine a first-order with a second-order signal. This is similar to the factor two that arises when the expression  $(a+b)^2$  is expanded to  $(a^2+2ab+b^2)$ .

**Simplification to memoryless systems** If it is assumed that both  $\mathbf{H}$  and  $\mathbf{F}$  represent memoryless systems, then the transfer functions involved in the equations (4.68) through (4.70) are independent of the frequency variable(s). If we represent the transfer functions of order one to three for the system  $\mathbf{H}$  by  $H_1$ ,  $H_2$  and  $H_3$  and the transfer functions for  $\mathbf{F}$  by  $F_1$ ,  $F_2$  and  $F_3$ , then we obtain for the overall transfer functions of order one to three

$$Q_1 = F_1 H_1 (4.71)$$

$$Q_2 = F_1 H_2 + F_2 H_1^2 (4.72)$$

$$Q_3 = F_3 H_1^3 + F_1 H_3 + 2F_2 H_1 H_2 (4.73)$$

The last term of  $Q_3$  may be negligible if a (linear) bandpass filter is placed between the systems H and F: this filter passes the wanted signal and suppresses out-of-band signals such as second-order harmonics and second-order intermodulation products. Hence, the effect of  $H_2$  is suppressed. Third-order intermodulation products, on the other hand, can be in-band signals which are not removed by such bandpass filter.

## 4.6.2 Application: a two-stage amplifier

Assume that the two cascaded stages are amplifying stages and assume for simplicity that capacitive and inductive effects are negligible. If, in addition, it is assumed that the two amplifiers are about identical, i.e.  $H_1 \approx F_1$ ,  $H_2 \approx F_2$ ,  $H_3 \approx F_3$ , then it is clear from equation (4.72) that the second-order nonlinearity of the second stage contributes a factor  $H_1$  more to the overall second-order nonlinear behavior than the first stage's second-order nonlinearity. The first two terms of equation (4.73) show that the difference in contribution to the overall third-order behavior between the two stages is a factor  $H_1^2$ . Roughly speaking, the largest contributions to the nonlinear distortion at the output of an amplifier originate from the circuit elements close to or at the output where signal swings are large. The considerations presented above are just a few interpretations of the expressions for a cascade of two nonlinear systems. The presented expressions can of course be applied in many other situations. An interesting application of the results obtained in this section is presented in the next subsection.

## 4.7 Pre-distortion and post-distortion using inverse systems

In the previous section equations for the cascade connection of two nonlinear systems have been derived. One could think of using these equations to find the requirements for a pth-order system.

H to compensate the nonlinearities of the nonlinear system F which is placed in cascade with H. The compensation would imply that the output of the overall system is a linear operation of the input, while the kernels from order two to order p are zero. In this case, the system H is the *inverse system* of F and vice versa. The above idea can — at least theoretically — be realized. When the system H is placed after the system F, then H accomplishes a post-distortion which compensates the nonlinear distortion of F. System H is then called the post-inverse of F. Conversely, when H is placed before F, one speaks about pre-distortion by the pre-inverse H. The general concepts are discussed in [Sche 80]. Some applications in analog integrated circuit design are given now. In [Sche 80] it is shown that the pth-order pre-inverse is identical to the pth-order post-inverse. Hence, the discussion can be limited to the post-inverse F of the system H.

#### 4.7.1 General expressions

The derivations of the expressions for the Volterra kernels of the pre-inverse are given in [Sche 80]. Only the Laplace transforms of the post-inverse **F** are discussed here.

The Laplace transform of  $\mathbf{F}_1$  is given by

$$F_1(s) = \frac{1}{H_1(s)} \tag{4.74}$$

Equation (4.74) reveals that the poles (zeros) of  $H_1(s)$  are the zeros (poles) of  $F_1(s)$ . Hence, if  $H_1(s)$  has zeros in the right half complex plane, then  $F_1(s)$  is the transfer function of an unstable linear system so that a stable pth-order inverse of  $\mathbf{H}$  does not exist. In the further discussion it is assumed that the linear operator  $\mathbf{F}_1$  satisfying equation (4.74) is stable.

The Laplace transform of  $\mathbf{F}_2$  is given by

$$F_2(s_1, s_2) = -\frac{H_2(s_1, s_2)}{H_1(s_1)H_1(s_2)H_1(s_1 + s_2)}$$
(4.75)

If **H** is a memoryless system then  $F_2(s_1, s_2)$  reduces to

$$F_2 = -\frac{H_2}{H_1^3} \tag{4.76}$$

The Laplace transform of the post-inverse of order three is already quite complicated [Sche 80]. Expressions for the Laplace transform of higher order, say order p, can be derived as well. If p would become infinite, then the obtained system consisting of all post-inverses is the inverse system of  $\mathbf{H}$ . A problem encountered in letting p go to infinity is that the corresponding Volterra series may converge only for a limited range of the input amplitude.

## 4.7.2 Example

As an example, consider the simplified expression of the collector current  $i_C$  of a bipolar transistor as a function of its base-emitter voltage  $v_{BE}$  (see equation (3.12)). From this equation the

AC collector current  $i_c$  can be found as a function of the AC base-emitter voltage  $v_{be}$ :

$$i_c = I_C \exp\left(\frac{v_{be}}{V_t}\right) - I_C \tag{4.77}$$

in which  $I_C$  is the quiescent collector current. For sufficiently small  $v_{be}$ , the exponential can be approximated well with a power series which is broken down after the first few terms:

$$i_c = g_m v_{be} \left[ 1 + \frac{1}{2} \frac{v_{be}}{V_t} + \frac{1}{6} \left( \frac{v_{be}}{V_t} \right)^2 + \dots \right]$$
 (4.78)

This relationship can be regarded as the description of a nonlinear memoryless system. The input variable is the voltage  $v_{be}$  and the output is the small-signal collector current  $i_c$ . The Laplace transform of the first-order kernel is found from equation (4.78). It is given by

$$H_1 = g_m \tag{4.79}$$

Since the system is memoryless, the dependence on the frequency variables has been omitted. The second-order kernel transform can also be found from equation (4.78) using the result of Section 4.3.1. We find

$$H_2 = \frac{g_m}{2V_t} \tag{4.80}$$

The above theory of the post-inverse can be applied to this system in order to find the inverse relation that expresses  $v_{be}$  as a function of  $i_c$ . From equation (4.74) we find for the first-order kernel of the inverse system

$$F_1 = \frac{1}{q_m} = \frac{V_t}{I_C} \tag{4.81}$$

and for the second-order kernel, using equation (4.75)

$$F_2 = \frac{V_t}{2I_C^2} {(4.82)}$$

In this simple case it is possible to write down an explicit expression for the inverse relationship of equation (4.77). This is found to be

$$v_{be} = V_t \ln \left( 1 + \frac{i_c}{I_C} \right) \tag{4.83}$$

Developing this relationship into a power series yields

$$v_{be} = \frac{V_t}{I_C} i_c - \frac{V_t}{2I_C^2} i_c^2 + \frac{V_t}{3I_C^3} i_c^3 \dots$$
 (4.84)

It is seen that the first- and the second-order coefficient of equation (4.84) correspond to  $F_1$  and  $F_2$  of equations (4.81) and (4.82).

## 4.7.3 Applications

In the next section we will see how feedback can be used to linearize the characteristic of an analog circuit. In cases where the use of feedback is too involved, pre- and post-distortion can be used. This occurs for example at high frequencies, where it is difficult to use feedback with a loop gain that is sufficiently high.

Consider a cascade of two nonlinear (sub)circuits where the first circuit performs a predistortion for the second circuit. Assume that the input of the first circuit is a current and the output is a voltage. Then the first circuit realizes a transimpedance. Since the input-output relationship of the second circuit must be the inverse of the input-output relationship of the first circuit (apart from a constant), the second circuit must realize a transadmittance or a transconductance. Similarly, when the first stage accomplishes a voltage to voltage or a current to current conversion, then of course the second stage must realize the same conversion.

The first examples of pre- and post-distortion are the two circuits of Figure 4.20. The left circuit is a current mirror: the applied input current is nonlinearly transformed to a voltage by transistor  $M_1$ , after which this voltage is transformed again to a current by transistor  $M_2$  that matches with  $M_1$ . In this way, a linear relationship is obtained between the output current  $i_{OUT}$  and the input current  $i_{IN}$ . Furthermore, the voltage over the linear resistor  $R_L$  depends linearly on the input current because  $R_L$  transforms the output current into a voltage in a linear way.

In the right circuit in Figure 4.20 a linear voltage-to-voltage conversion is realized by the cascade of a nonlinear voltage-to-current conversion (transistor  $M_1$ ) and the inverse current-to-voltage conversion of the output impedance seen at the source of transistor  $M_2$ .

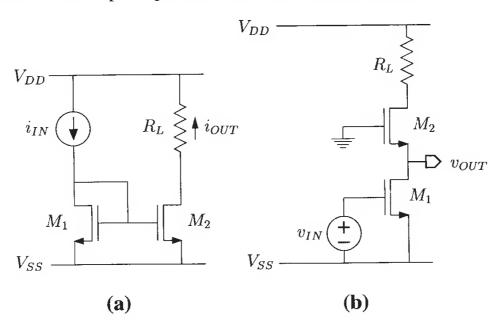


Figure 4.20: Application of the pre- or post-distortion principle: (a) cascade of a current-to-voltage converter followed by the inverse voltage-to-current converter. (b) The reverse situation of (a).

It must be noticed that the pre- and post-distortion in the two above examples are perfect if the

transconductances are the only circuit elements taken into account in the transistors' equivalent circuit. Parasitic effects due to the presence of other circuit elements partially destroy the pre- or post-distortion. Their effect can be calculated with the computation method that is explained in Section 5.2.

Another example is the pre-distortion circuit in a Gilbert multiplier [Gilb 68], depicted in Figure 4.21. In Section 4.5.1 we computed that the input-output relationship of a differential pair at low frequencies is given by a hyperbolic tangent relationship. One can compute [Gilb 68, Gray 93] that the pre-distortion circuit consisting of transistors  $Q_{1A}$  and  $Q_{1B}$  accomplishes an inverse hyperbolic tangent relationship between the differential input current and the voltage  $\Delta V$  that is applied to the differential pair. The latter circuit maps a differential voltage to a differential current with a hyperbolic tangent relationship. When the order of both circuits is reversed, as shown in Figure 4.22, a linear voltage-to-voltage conversion is realized. This conversion is linear as long as the base current is neglected.

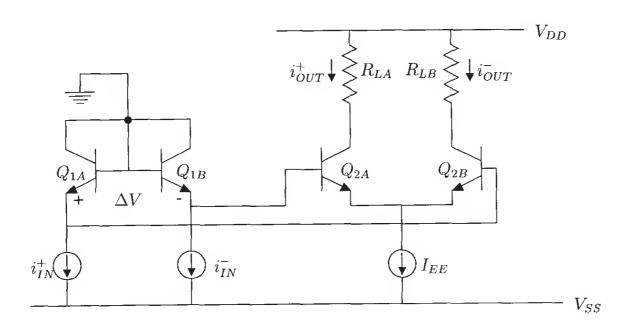


Figure 4.21: Linear current-to-current conversion in a Gilbert multiplier with pre-distortion.

In the derivations and the examples above it has been assumed that the gain of the overall circuit that consists of the cascade of the system and its pre- or post-inverse, is equal to one. However, a larger gain together with the benefits of pre- or post-distortion can be accomplished. If the Volterra kernels of the second block are all multiplied by a constant a, then from the general expressions (4.68) through (4.70) for a cascade connection, it is seen that this factor a is common to all terms and so the overall kernels are all multiplied by a. This property is used for example in current mirrors to obtain a linear current gain, but it can be generalized as well to other circuits that use pre- or post-distortion.

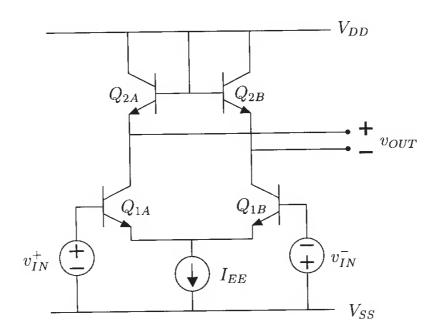


Figure 4.22: Linear voltage-to-voltage conversion using post-distortion.

## 4.8 Linear and nonlinear feedback

Linear feedback is widely used in analog circuits because of its important benefits such as stabilization of the gain of an amplifier against parameter changes, the increase of the bandwidth, the ability for a designer to modify the terminal impedances of a circuit in almost any fashion, .... Yet another benefit, which will be studied in this section, is the reduction of the nonlinear distortion. The latter benefit has been studied for example in [Nar 70]. Nevertheless, designers often know this property qualitatively but not quantitatively, especially at high frequencies.

In this section we will also consider nonlinear feedback. Applications of nonlinear feedback are seldom found in the analog design community. However, as will be shown in this section, nonlinear feedback can be applied usefully as well.

Figure 4.23 depicts the most general feedback structure that is studied here. In most analog integrated circuit applications, the operator **H** represents an amplifier, denoted as the *basic amplifier*, whereas operator **F** corresponds to the *feedback network*, which is usually passive. In this section the nonlinear transfer functions for the overall system, denoted as the *feedback system* or the *closed-loop system* and represented by **Q**, are computed and interpreted. First, both the basic amplifier and the feedback network are assumed to be nonlinear. From the general expressions some special cases will be derived.

It is important to note that the basic amplifier can either be a voltage amplifier, a current amplifier, a transconductance or a transresistance. The same applies for the feedback network. For example, when the input of Figure 4.23 is a voltage and the output is a current, then the basic amplifier is a transconductance, whereas the feedback network, that measures the output current and feeds back a voltage, is a transresistance.

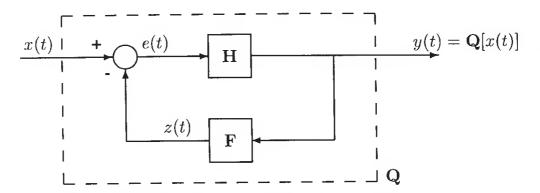


Figure 4.23: A general nonlinear feedback system.

#### 4.8.1 Nonlinear feedback systems

The nonlinear transfer functions from order one to three are first calculated for both **H** and **F** of Figure 4.23 being nonlinear. In [Sche 80] expressions in operator notation are presented for the first three Volterra kernels of **Q**. A translation of these equations to the frequency domain is quite involved. The results are presented below.

The first-order transfer function of the closed-loop system is given by

$$Q_1(s_1) = H_1(s_1)R(s_1) (4.85)$$

Hereby R(s) is a shorthand notation for

$$R(s) = \frac{1}{1 + H_1(s)F_1(s)} \tag{4.86}$$

Equation (4.85) shows that the linear gain of the basic amplifier is reduced by a factor R(s) whose magnitude is usually much smaller than one. It is assumed that R(s) and  $Q_1(s)$  do not have poles in the right half of the s-plane.

Besides its interpretation as a **gain reduction factor**, the function R(s) can — after multiplication with X(s), the Laplace transform of x(t) — also be regarded as the Laplace transform of the signal e(t) in Figure 4.23. Both interpretations are useful for the interpretation of higher-order transfer functions.

The second-order transfer function  $Q_2(s_1, s_2)$  consists of two parts that share some common factors:

$$Q_2(s_1, s_2) = R(s_1)R(s_2) \left[ H_2(s_1, s_2) - H_1(s_1)H_1(s_2)F_2(s_1, s_2)H_1(s_1 + s_2) \right] R(s_1 + s_2)$$
(4.87)

In this equation the common factors are placed outside the square brackets. The first term between the brackets is due to the second-order nonlinearity of the basic amplifier only, while the second term is due to the nonlinearity of the feedback network only. For their interpretation one can start from two fundamental frequencies before block  $\mathbf{H}$ , whose Laplace transforms are  $R(s_1)X(s_1)$  and  $R(s_2)X(s_2)$ , respectively. The second-order operator  $\mathbf{H}_2$  acts upon these frequencies to produce a second-order output, corresponding to the first term between the square

brackets. Apart from this, the two signals  $R(s_1)X(s_1)$  and  $R(s_2)X(s_2)$  are both linearly amplified by  $\mathbf{H}_1$  after which they are supplied to  $\mathbf{F}_2$  which produces a second-order signal. This signal is then linearly amplified at the frequency  $s_1+s_2$  by  $\mathbf{H}_1$ , yielding the second term between the square brackets of equation (4.87). The minus sign of this term corresponds to the minus sign in Figure 4.23. Finally, both second-order signals at the output of the amplifier are reduced by a factor  $R(s_1+s_2)$  as indicated by the common factor at the end of the expression.

The expression for the third-order transfer function is more involved and has been split in three parts that again share some common factors:

$$Q_{3}(s_{1}, s_{2}, s_{3}) = \prod_{i=1}^{3} R(s_{i}) \begin{bmatrix} H_{3}(s_{1}, s_{2}, s_{3}) - 2H_{2}(s_{1}, s_{2})F_{1}(s_{1} + s_{2})R(s_{1} + s_{2})H_{2}(s_{3}, s_{1} + s_{2}) \\ \dots \\ + \prod_{i=1}^{3} H_{1}(s_{i}) \Big( -F_{3}(s_{1}, s_{2}, s_{3}) \\ + 2F_{2}(s_{1}, s_{2})H_{1}(s_{1} + s_{2})R(s_{1} + s_{2})F_{2}(s_{3}, s_{1} + s_{2}) \Big) \cdot H_{1}(s_{1} + s_{2} + s_{3}) \\ \dots \\ - 2H_{2}(s_{1}, s_{2})R(s_{1} + s_{2})H_{1}(s_{3})F_{2}(s_{1}, s_{2} + s_{3})H_{1}(s_{1} + s_{2} + s_{3}) \\ - 2H_{1}(s_{1})H_{1}(s_{2})F_{2}(s_{1}, s_{2})R(s_{1} + s_{2}) \cdot H_{2}(s_{3}, s_{1} + s_{2}) \Big] \cdot R(s_{1} + s_{2} + s_{3})$$

$$(4.88)$$

The three parts between the square brackets have been separated by a dotted line for clarity and the common factors are placed outside the square brackets. The expression is formulated in a non-symmetrized form. Symmetrization makes the expression even more complicated and hence does not contribute to an easier interpretation.

The first part between the square brackets of equation (4.88) consists of two terms, which correspond to the nonlinearities of the basic amplifier only: the first term is due to the third-order operator  $\mathbf{H}_3$  that acts upon three fundamental frequencies  $R(s_1)X(s_1)$ ,  $R(s_2)X(s_2)$  and  $R(s_3)X(s_3)$ . The second term is caused by the interaction of  $\mathbf{H}_2$  with itself: this operator first combines two signals at  $s_1$  and  $s_2$ . The signal at the output of  $\mathbf{H}_2$  is then fed back and then again presented to  $\mathbf{H}_2$ . This corresponds to a multiplication with  $(-F_1(s_1+s_2)R_1(s_1+s_2))$ . Finally, this signal is fed into  $\mathbf{H}_2$  together with a third first-order input term at  $s_3$ . The factor two is caused by the last operation by  $\mathbf{H}_2$  which can combine the first and the second-order frequency in two ways<sup>2</sup>. Finally, both terms are divided by the reduction factor evaluated at  $s_1 + s_2 + s_3$ . This reduction is the same for the other contributions discussed below and is therefore placed at the end of the complete expression.

The second part between the square brackets is due to nonlinearities in the feedback network only. It consists of two terms that share some common factors which are placed outside the

<sup>&</sup>lt;sup>2</sup>This factor two is similar to the factor two that arises when the expression  $(a+b)^2$  is expanded to  $(a^2+2ab+b^2)$ .

rounded brackets. The first term is caused by the third-order operator  $\mathbf{F}_3$  of the feedback network that acts upon three fundamental frequencies at the output of the basic amplifier. The minus sign corresponds to the minus sign in Figure 4.23. The second term is caused by the interaction of  $\mathbf{F}_2$  with itself: two fundamental frequencies at the basic amplifier's output are fed into  $\mathbf{F}_2$ , which produces a second-order output. This signal is linearly amplified by  $\mathbf{H}_1$  at  $s_1 + s_2$ , reduced by  $R(s_1 + s_2)$  and is then, together with a third fundamental frequency at the basic amplifier's output, acted upon by  $\mathbf{F}_2$ . Since the feedback loop has been walked through twice, the two minus signs of Figure 4.23 cancel each other.

The third part of  $Q_3(s_1, s_2, s_3)$  is caused by the interaction of the second-order nonlinearity of the basic amplifier and the feedback network. For the first term two fundamental frequencies are combined by  $\mathbf{H}_2$  to a second-order signal. This signal is reduced by  $R(s_1 + s_2)$  and is then, together with a first-order output of the basic amplifier, fed to  $\mathbf{F}_2$ , which produces a third-order output. This output is then multiplied by -1 and linearly amplified by  $\mathbf{H}_1$ . The last term is produced by a second-order output at the output of  $\mathbf{F}_2$  which is taken together with a third fundamental input frequency by  $\mathbf{H}_2$ .

#### 4.8.2 Feedback with a large loop gain

In many applications the loop gain T(s) which is given by

$$T(s) = F_1(s)H_1(s) (4.89)$$

has a magnitude much larger than one. Then it is seen from equation (4.86) that the reduction factor R(s) is about equal to 1/T(s). Under this assumption, the transfer functions given in equations (4.85), (4.87) and (4.88) reduce to

$$Q_{1}(s_{1}) = \frac{1}{F_{1}(s_{1})}$$

$$Q_{2}(s_{1}, s_{2}) = \frac{H_{2}(s_{1}, s_{2})}{T(s_{1})T(s_{2})T(s_{1} + s_{2})} - \frac{F_{2}(s_{1}, s_{2})}{F_{1}(s_{1})F_{1}(s_{2})F_{1}(s_{1} + s_{2})}$$

$$Q_{3}(s_{1}, s_{2}, s_{3}) = \frac{1}{T(s_{1})T(s_{2})T(s_{3})} \frac{1}{T(s_{1})T(s_{1})T(s_{2})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s_{1})T(s_{1})T(s_{1})} \frac{1}{T(s_{1})T(s$$

$$+\frac{-2H_2(s_1,s_2)F_2(s_1,s_2+s_3)}{T(s_1+s_2)T(s_1)T(s_2)F_1(s_3)F_1(s_1+s_2+s_3)}$$

$$+\frac{-2F_2(s_1,s_2)H_2(s_3,s_1+s_2)}{T(s_1+s_2)T(s_3)T(s_1+s_2+s_3)F_1(s_1)F_1(s_2)}$$
(4.92)

Equation (4.90) shows the well-known result that in the presence of a large loop gain the closed-loop gain of the linearized circuit is the inverse of the first-order transfer function of the feedback network.

It is seen both in equation (4.91) and in equation (4.92) that the second- and third-order transfer functions contain terms that have T(s) in the denominator. This means that these terms can be suppressed by increasing the loop gain. It is seen that these terms correspond to the nonlinearities in the basic amplifier only. Both the nonlinear transfer function of order two and three contain terms that are not suppressed by the loop gain. These terms correspond to the nonlinearities in the feedback network. This means that a large loop gain of the linearized feedback can suppress the nonlinearities in the basic amplifier but it cannot eliminate the nonlinearities in the feedback network. In other words, the nonlinearities in the feedback network determine the lower limit of the suppression of the harmonics or intermodulation products by a large loop gain. Therefore, for very low-distortion applications it is important to use a feedback network that is as linear as possible.

Similar to the nonlinear effects of the basic amplifier only, the terms caused by the interaction of the nonlinearities of the basic amplifier and the feedback network are lowered when the loop gain is large. However, in equation (4.92) it is seen that these interaction terms are only divided by three loop gains, whereas the terms caused by the nonlinearities of the basic amplifier only are divided by four loop gains. This means that the contribution to the distortion from the interaction terms is only kept small because the signals are lowered by the loop gain and no additional suppression occurs anymore.

#### 4.8.3 Linear feedback

We now consider the case where the feedback network is perfectly linear. The use of linear feedback is the most widely used technique to reduce nonlinear distortion. Therefore, this form of feedback is studied extensively. First, the general expressions are derived. Next, we simplify the expressions to feedback with a large loop gain and to memoryless systems. Finally, a practical example is considered.

In many applications the feedback network represented by F in Figure 4.23 is a passive circuit which can be considered as being linear. In this case,  $F_2(s_1, s_2) = 0$  and  $F_3(s_1, s_2, s_3) = 0$ . Then the expressions for the second and third-order transfer function reduce to the first part of equations (4.87) and (4.88), respectively:

$$Q_1(s_1) = H_1(s_1)R(s_1) (4.93)$$

$$Q_2(s_1, s_2) = R(s_1)R(s_2)R(s_1 + s_2)H_2(s_1, s_2)$$
(4.94)

$$Q_{3}(s_{1}, s_{2}, s_{3}) = R(s_{1})R(s_{2})R(s_{3})[H_{3}(s_{1}, s_{2}, s_{3}) - 2H_{2}(s_{1}, s_{2})F_{1}(s_{1} + s_{2})R(s_{1} + s_{2})H_{2}(s_{3}, s_{1} + s_{2})] \cdot R(s_{1} + s_{2} + s_{3})$$

$$(4.95)$$

The expression of the linear gain  $Q_1(s_1)$  is of course the same as in the linear case. Further it is seen that the second-order transfer function of the basic amplifier is reduced by a factor  $R(s_1)R(s_2)R(s_1+s_2)$ . If, however, the reduction of the linear gain is compensated by an increase of the input level with a factor R(s) ( $s = s_1$  or  $s_2$ ), then the second-order transfer function is only reduced by a factor  $R(s_1 + s_2)$ .

Similarly, it is found from equation (4.95) that, after correction of the input level, the third-order nonlinearity is suppressed with a factor  $R(s_1 + s_2 + s_3)$ . It is seen that the second-order nonlinearity of the basic amplifier also contributes to the third-order kernel of  $\mathbf{Q}$ . This contribution is also suppressed by the loop gain and it has a sign that is opposite to the sign of the contribution of the third-order nonlinearity of the basic amplifier. to the sign of the contribution of second-order nonlinearity. It is seen that the third-order kernel can be made zero by adjusting the loop gain such that the contribution of the second-order nonlinearity cancels with the contribution of the third-order nonlinearity. This can be seldom exploited in practice [Lot 68, Tho 68], since the value of the loop gain must be well controlled and it must be relatively small.

Linear feedback, however, does not always decrease the nonlinear distortion. In some circuits [Khour 87] positive feedback is used with a loop gain that is smaller than one. In this case, the absolute value of the reduction factor R(s) is larger than one. Hence, the nonlinear distortion can increase considerably in such circuits.

#### 4.8.3.1 Simplification to memoryless systems

For memoryless systems the expressions for the kernel transforms obtained above (equations (4.93 through (4.95)) can be further simplified. In this case, the involved kernel transforms are independent of the Laplace variables  $s_1$ ,  $s_2$  and  $s_3$ . For the sake of clarity these variables can be omitted as arguments. Then the kernel transforms become

$$Q_1 = H_1 R \tag{4.96}$$

$$Q_2 = H_2 R^3 (4.97)$$

$$Q_3 = [H_3 - 2H_2^2 F_1 R] R^4 (4.98)$$

Assume now further that the loop gain is very large. In this case R reduces to  $1/T = 1/(F_1H_1)$ . Then one obtains

$$Q_1 = \frac{H_1}{T} = \frac{1}{F_1} \tag{4.99}$$

$$Q_2 = \frac{H_2}{T^3} \tag{4.100}$$

$$Q_3 = \left[ H_3 - \frac{2H_2^2}{H_1} \right] \frac{1}{T^4} \tag{4.101}$$

At this moment it is interesting to compare the harmonic distortion of the feedback circuit to the distortion of the basic amplifier in an open loop configuration (without feedback). From

equations (2.13) and (2.14) we recall that the harmonic distortion of the basic amplifier that is modeled with  $H_1$ ,  $H_2$  and  $H_3$  is given by

$$HD_{2 \text{ open loop}} = \frac{A_1}{2} \frac{H_2}{H_1} \tag{4.102}$$

$$HD_{2 \text{ open loop}} = \frac{A_1}{2} \frac{H_2}{H_1}$$
 (4.102)  
 $HD_{3 \text{ open loop}} = \frac{A_1^2}{4} \frac{H_3}{H_1}$  (4.103)

where  $A_1$  is the input amplitude. In the closed loop situation we already know that the gain is reduced by a factor 1/T. Compensating this reduction by applying a signal with an amplitude of  $TA_1$  yields for  $HD_2$  and  $HD_3$  in closed loop

$$HD_{2 \, \text{closed loop}} = \frac{A_1}{2} \frac{1}{T} \frac{H_2}{H_1}$$
 (4.104)

$$HD_{3 \text{ closed loop}} = \frac{A_1^2}{4} \frac{1}{T} \left| \frac{H_3}{H_1} - \frac{2H_2^2}{H_1^2} \right| \tag{4.105}$$

It is seen that the second harmonic distortion is reduced by a factor T. The third harmonic distortion consists of two contributions: apart from the contribution of the third-order nonlinearity of the basic amplifier which is suppressed by a factor T, there is also a contribution of the secondorder nonlinearity with an opposite sign. The above expressions will be applied to a simple example in Section 4.8.3.2.

Qualitative explanation The suppression of harmonics by applying feedback with a large loop gain can be explained qualitatively as follows [Gray 93]. Assume that the input-output relationship of the basic amplifier is given by the characteristic of Figure 4.24. This characteristic is seen

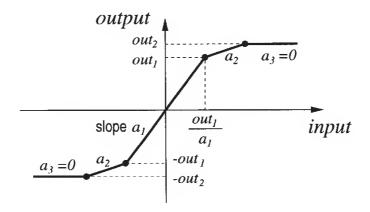


Figure 4.24: Transfer characteristic of a basic amplifier.

to be composed of three parts, each having a different slope. In reality, this artificial separation into these regions is less abrupt, but this is not important here. The part around the origin has a slope  $a_1$  that corresponds to the gain of the linearized amplifier. For larger signals the transfer characteristic has a slope  $a_2$  which is somewhat smaller than  $a_1$ . For very large inputs the amplifier saturates, which is reflected by a zero slope.

Assume now that the gain of the amplifier is very large but the absolute value is badly known and may depend for example on temperature. In order to stabilize the gain, feedback with a large loop gain is applied. Neglecting frequency effects it is found from equation (4.99) that the gain of the linearized closed-loop configuration only depends on the characteristics of the feedback network:

$$gain = \frac{a_1}{1+T} \approx \frac{1}{F_1} \tag{4.106}$$

When, at large signal levels, the gain of the basic amplifier slightly decreases due to a change of the slope of the transfer characteristic, then equation (4.106) still holds, since the loop gain is still large. In other words, the closed loop gain is less dependent on the input amplitude, resulting in less nonlinear distortion. However, for very large input levels, the amplifier saturates. At that moment, the loop gain decreases dramatically. As a result, it cannot stabilize the gain anymore.

The transfer of the closed-loop amplifier is depicted in Figure 4.25. It is seen that for small input levels the slope is lower compared to the open loop case. This corresponds to a lower gain. Also, the slope for output levels higher than  $out_1$ , is closer to the slope for small input signals. This corresponds to a linearization.

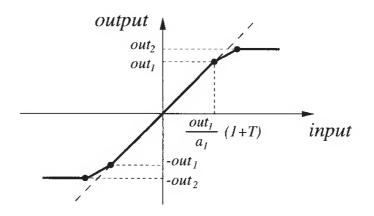


Figure 4.25: Transfer characteristic of a feedback configuration with the basic amplifier with the characteristic of Figure 4.25.

#### 4.8.3.2 Application: emitter degeneration

Figure 4.26 shows an amplifier consisting of one bipolar transistor with emitter degeneration. The circuit is driven by a voltage source  $v_{IN}$ . The output of interest is the current through the load resistance  $R_L$ . The emitter degeneration can be considered as a series-series feedback configuration [Gray 93]. The loop gain of this feedback is equal to  $g_m R_E$ , where  $g_m$  is the transconductance of transistor  $Q_1$ .

For the analysis here, only the collector current of transistor  $Q_1$  is taken into account. The collector current is modeled as a nonlinear transconductance. Its linear behavior is represented by the transconductance  $g_m$  and the second- and third-order nonlinearity by the coefficients  $K_{2g_m}$ 

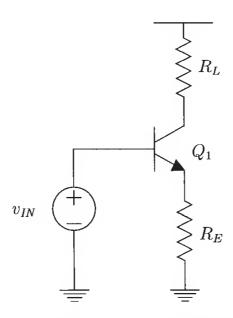


Figure 4.26: A single-transistor amplifier with emitter degeneration.

and  $K_{3g_m}$ , respectively, as explained in Section 3.2.1. Using a simple exponential model for the collector current (equation (3.12)) it was found there that  $g_m = I_C/V_t$ ,  $K_{2g_m} = g_m/(2V_t)$  and  $K_{3g_m} = g_m/(6V_t^2)$ .

For the sake of simplicity the loading of the feedback network, which is nothing more than resistor  $R_E$ , is neglected. Hence the feedback configuration can be represented as in Figure 4.27. In this circuit it is seen that the output current, in this case the collector current through  $Q_1$ , is measured and fed back by means of a voltage-controlled voltage source. From this figure it is

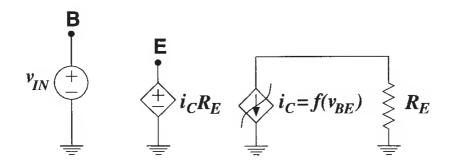


Figure 4.27: Equivalent circuit of the amplifier of Figure 4.26.

clear that the basic amplifier of the general feedback configuration of Figure 4.23 can be identified with the nonlinear transconductance and the feedback network with the (linear) current-controlled voltage source.

Recalling the discussion from Sections 4.3.1 and 4.3.2, we can associate  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$ 

to the first-, second- and third-order kernel transform of the basic amplifier:

$$H_1 = g_m \tag{4.107}$$

$$H_2 = K_{2q_m} (4.108)$$

$$H_3 = K_{3q_m} (4.109)$$

The transfer function of the feedback network, corresponding to  $F_1$  in the equations (4.96) through (4.98) is given by

$$F_1 = R_E \tag{4.110}$$

Hence we find for the reduction factor R

$$R = \frac{1}{1 + q_m R_E} \tag{4.111}$$

When the input voltage source is a sinusoid,  $v_{in} = A_1 \sin(\omega_1 t)$ , then the kernel transforms of the output current can be found using equations (4.96) through (4.98). This yields

$$I_{out_{1,0}} = \frac{g_m}{1 + q_m R_E} \tag{4.112}$$

$$I_{out_{2,0}} = \frac{g_m}{2V_t \left(1 + q_m R_E\right)^3} \tag{4.113}$$

$$I_{out_{3,0}} = \frac{1}{\left(1 + g_m R_E\right)^4} \cdot \frac{g_m}{2V_t^2} \left[ \frac{1}{3} - \frac{g_m R_E}{1 + g_m R_E} \right]$$
(4.114)

It is seen that the third-order kernel becomes zero for  $g_m R_E = 0.5$ . This is a very small value in practice. Assume for example that the DC collector current of  $Q_1$  is 1mA. Then  $g_m \approx 40mA/V$ . A loop gain of 0.5 then corresponds to an emitter resistance of  $12.5\Omega$ . This value is of the same order of magnitude as the parasitic emitter resistance of a bipolar transistor.

The kernel transforms of order one to three of the output voltage can be easily found by multiplying the kernels of the collector current with  $-R_L$ .

Consider now the harmonic distortion of the output current. The values that will be obtained will also correspond to the harmonic distortion figures for the output voltage since the kernel of order one to three of the output current and the output voltage differ by a constant factor, the vanishes when the ratio of two harmonics is taken. From equations (4.112) and (4.114) we find

$$HD_2 = \frac{A_1}{2} \frac{1}{2V_t \left(1 + g_m R_E\right)^2} \tag{4.115}$$

$$HD_3 = \frac{A_1^2}{4} \frac{1}{\left(1 + g_m R_E\right)^3} \frac{1}{2V_t^2} \left| \frac{1}{3} - \frac{g_m R_E}{1 + g_m R_E} \right| \tag{4.116}$$

Assume now that  $g_m R_E \gg 1$ . Then the harmonic distortion figures reduce to

$$HD_2 \approx \frac{A_1}{4V_t \left(q_m R_E\right)^2} \tag{4.117}$$

$$HD_3 pprox rac{A_1^2}{12V_t^2 \left(g_m R_E\right)^3}$$
 (4.118)

Clearly, the harmonic distortion can be reduced by increasing  $g_m$  or  $R_E$ , corresponding to a higher loop gain.

### 4.8.4 Nonlinearities in the feedback network

We now consider a feedback configuration where the basic amplifier is linear and the feedback network is nonlinear. It has already been mentioned that the effect of the nonlinearities in the feedback network is not reduced by the loop gain.

An interesting but still little exploited effect of nonlinearities in the feedback circuit is that the nonlinearities at the output of a closed-loop system with a large loop gain are exactly the inverse of the nonlinearities of the feedback circuit itself. The Laplace transform of the inverse of a nonlinear system, shown in equation (4.75) indeed matches with the second part of equation (4.91) which is the second-order transfer function of the closed-loop system caused by the second-order nonlinearity of the feedback circuit only. This similarity can also be proven for higher orders. This means that in the absence of nonlinearities of the basic amplifier, a feedback system yields the inverse of the operator that describes the feedback network. This is generally known for linear systems but it can be extended to weakly nonlinear systems.

## 4.8.5 Loading effect of a nonlinear feedback network

In most practical feedback configurations the feedback circuit forms a load for the basic amplifier both at its input and output. The nonlinear feedback formulas derived above, do not include loading effects.

In linear circuits the loading of the feedback network can be included with the basic amplifier so that ideal feedback equations again apply [Gray 93]. For two-port feedback circuits, this is possible through the use of linear two-port parameters such as y, z, g and h parameters, which establish a galvanic separation between the two ports.

This splitting can in general not be performed for nonlinear elements. Consider for example the nonlinear shunt-shunt feedback configuration shown in Figure 4.28. The current through the

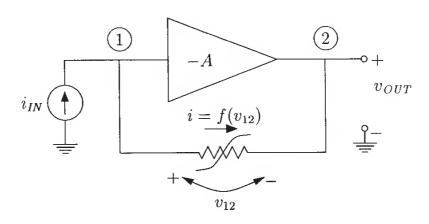


Figure 4.28: A nonlinear shunt-shunt feedback configuration.

nonlinear conductance is a function of the voltage difference  $v_{12} = v_1 - v_2$ . Using the power series expansion defined in Section 3.2, this current i can be written as

$$i = g_1 \cdot v_{12} + K_{2g_1} \cdot v_{12}^2 + K_{3g_1} \cdot v_{12}^3 + \dots$$

$$= g_1 \cdot v_1 - g_1 \cdot v_2 + K_{2g_1} \cdot v_1^2 + K_{2g_1} \cdot v_2^2 - 2K_{2g_1} \cdot v_1 \cdot v_2$$

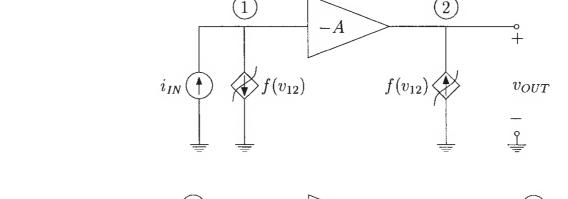
$$(4.119)$$

$$+ K_{3g_1} \cdot v_1^3 + K_{3g_1} \cdot v_2^3 - 3K_{3g_1} \cdot v_1^2 \cdot v_2 + 3K_{3g_1} \cdot v_1 \cdot v_2^2$$

$$(4.120)$$

$$= f_1(v_1) + f_2(v_2) + f_3(v_1, v_2)$$
(4.121)

In this way, the one-dimensional nonlinearity has been written as a two-dimensional nonlinearity. The functions  $f_1$  and  $f_2$  are functions of one variable. The function  $f_3$  corresponds to those terms in the series of equation (4.120) that contain both  $v_1$  and  $v_2$ . Using this reformulation, the circuit of Figure 4.28 has been redrawn in Figure 4.29. In this way, a galvanic separation has been



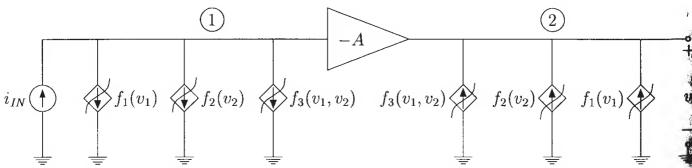


Figure 4.29: Equivalent representation of the shunt-shunt feedback circuit of Figure 4.28.

established. The two nonlinearities corresponding to the cross-terms of equation (4.120) remain controlled by the input and the output voltage. If these cross-terms are not present, such as for linear elements, or when they can be neglected, then the remaining elements can be merged with the basic amplifier or with the feedback circuit. Indeed, the nonlinearity  $f_1(v_1)$  at the input can be associated with an extra admittance which is added to the input admittance of the basic amplifier. The nonlinearity  $f_2(v_2)$  at the input is then the nonlinear feedback network which does not load the amplifier.

At the output, the nonlinearity  $f_1(v_1)$  can be considered as an extra forward signal path in the basic amplifier. The gain along this signal path is small compared to the gain of the basic

amplifier if the feedback network is passive. Nonlinearity  $f_2(v_2)$  at the output adds to the output admittance of the basic amplifier.

If the cross-terms in equation (4.120) would not be present, then the formulas derived in the previous sections can be applied using the new basic amplifier and the new feedback network. However, the cross-terms can be as large as the other terms, which means that the feedback formulas derived in the previous sections are no longer exact when the loading of the basic amplifier by the feedback network is considerable.

## 4.8.6 Operational amplifier in a linear feedback configuration

We consider again the inverting amplifier of Figure 1.2 which is redrawn here for convenience in Figure 4.30. The amplifier is loaded at its output with a resistance  $R_L$ . It is assumed that the operational amplifier is nonlinear and it has a finite bandwidth. We will now derive the second and third harmonic distortion for the complete amplifier.

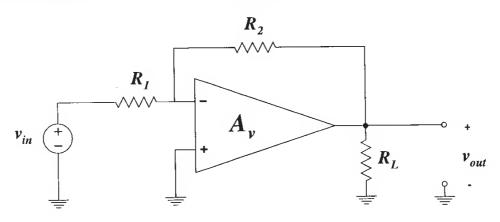


Figure 4.30: An inverting amplifier.

Model of the operational amplifier First we will set up a simplified model of the operational amplifier. It is assumed that the amplifier is an inverting two-stage amplifier as given in Figure 4.31. For the frequency behavior only the dominant pole of the opamp is taken into account. The pole of the opamp is at very low frequencies and it is determined by the capacitor  $C_c$  at the output of the first stage. The first stage is modeled as a differential amplifier with gain  $A_{v10}$ . The second stage is modeled as a common source amplifier that is loaded with a resistor  $R_{Lint}$ . The output of the second stage is fed into a buffer with a voltage gain of 1.

The pole of the opamp is taken into account in the frequency dependence of the first stage amplification. In this way, the first-stage gain, which is the voltage gain from the input to the internal node 1, is given by

$$A_{v1}(s) = \frac{V_1}{V_{id}} = \frac{A_{v10}}{1 + \frac{s}{p_d}}$$
(4.122)

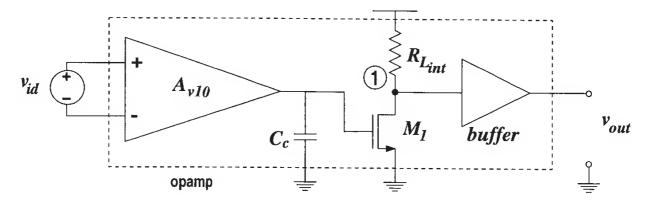


Figure 4.31: Equivalent circuit of a two-stage operational amplifier.

where  $p_d$  represents the dominant pole of the opamp in radians per second and  $A_{v10}$  is the low-frequency value of the gain of the first stage.

The only nonlinearity that is taken into account in this example is the nonlinearity of the drain current of transistor  $M_1$ . This drain current is described by the transconductance of  $M_1$ , namely  $g_m$ , and the second- and third-order nonlinearity coefficients  $K_{2g_m}$  and  $K_{3g_m}$ .

Nonlinearities in the first stage are neglected. This is justified by the fact that in a cascade connection of two amplifying blocks the nonlinearities of the second stage are dominant for the second- and third-order Volterra kernel of the cascade connection (see Section 4.6). This will also be illustrated in Chapter 8 where a detailed computation of the harmonics of a Miller-compensated operational amplifier is performed, including the nonlinearities of all transistors.

We will now analyze the nonlinear AC behavior of the amplifier in the frequency domain. The total gain of the linearized operational amplifier is the product of the gain of the two stages:

$$A_v(s) = \frac{V_{out}}{V_{id}} = A_{v1}(s) \left( -g_m R_{L_{int}} \right) = -\frac{A_{v10} g_m R_{L_{int}}}{1 + \frac{s}{p_d}}$$
(4.123)

For the rest of this section it is assumed that the frequency of the signals that are applied to the circuit have a frequency well above  $|p_d|/(2\pi)$  but still much lower than the gain-bandwidth product of the opamp, which is given by

$$GBW = (A_{v10}g_m R_{L_{int}}|p_d|)/(2\pi)$$
(4.124)

Hence equation (4.123) can be rewritten as

$$A_v(s) \approx \frac{-2\pi \, GBW}{s} \tag{4.125}$$

Finally, the input impedance of the opamp is assumed to be infinite whereas the output impedance is zero.

Volterra kernels of the basic amplifier and the feedback network First we will redraw the circuit of Figure 4.30 such that the basic amplifier and the feedback network can be easily identified. To this purpose, we first represent the input voltage source and resistor  $R_1$  by a Norton equivalent, as shown in Figure 4.32. The input is now a current source  $i_{in}(t)$  with a value of

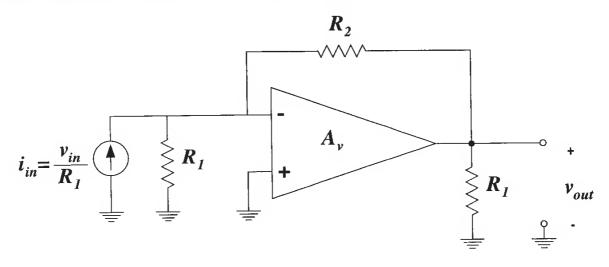


Figure 4.32: Inverting amplifier with  $v_{in}$  and  $R_1$  replaced by their Norton equivalent circuit.

$$i_{in} = \frac{v_{in}}{R_1}$$
 (4.126)

Next, resistor  $R_2$  is represented by a two-port that is described with y-parameters [Gray 93]. The y-parameters of a general two-port, shown in Figure 4.33, relate the currents that enter the two-port to the voltages over the two ports as follows:

$$i_1 = y_{11}v_1 \bigg|_{v_2 = 0} + y_{12}v_2 \bigg|_{v_1 = 0}$$
(4.127)

$$i_2 = y_{21}v_1 \bigg|_{v_2 = 0} + y_{22}v_2 \bigg|_{v_1 = 0}$$
 (4.128)

The reason for the use of a two-port representation of resistor  $R_2$  is that in this way the load-

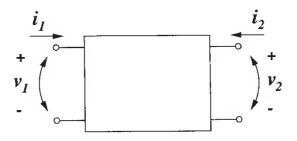


Figure 4.33: General two-port representation of a circuit.

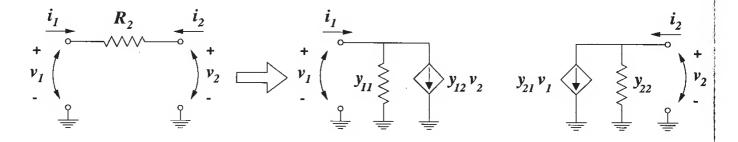


Figure 4.34: Two-port representation of resistor  $R_2$  from Figure 4.32.

ing of feedback resistor at the input and output of the basic amplifier can be taken into account [Gray 93]. The two-port representation of  $R_2$  is shown in Figure 4.34. From this figure we easily find

$$y_{11} = \frac{1}{R_2} \tag{4.129}$$

$$y_{12} = -\frac{1}{R_2} \tag{4.130}$$

$$y_{21} = -\frac{1}{R_2} \tag{4.131}$$

$$y_{22} = \frac{1}{R_2} \tag{4.132}$$

With this two-port representation the basic amplifier and the feedback network can now be identified as shown in Figure 4.35.

The basic amplifier contains the operational amplifier with the resistors  $R_1$  and  $R_2$  at its input and  $R_2$  and  $R_L$  at its output. The resistors with value  $R_2$  at the input and output correspond to the inverse of the y parameters  $y_{11}$  and  $y_{22}$  from Figure 4.34. It is seen that the basic amplifier is a transresistance amplifier: the input signal is a current and the output is a voltage.

It should be noticed that the voltage-controlled current source  $y_{21}v_1$  of the y-parameter representation of  $R_2$  has been omitted: this current source is shorted by the voltage-controlled voltage source that represents the opamp. If the output impedance of the opamp would not be zero, then this voltage-controlled current source would play a (little) role.

The feedback network is seen to consist of a voltage-controlled current source. This corresponds to the controlled source  $y_{12}v_2$  of Figure 4.34.

The feedback configuration that is shown in Figure 4.35 can now be identified with the general representation of a feedback system as shown in Figure 4.23. Then the kernel transforms of the operators H and F that represent the basic amplifier and the feedback network, respectively, will be determined, and finally the formulas (4.93) through (4.95) will be applied to find the kernel transforms of the overall feedback system Q. To this purpose, the configuration of Figure 4.35 is redrawn in a block diagram as shown in Figure 4.36. In this figure the voltage-controlled voltage source of Figure 4.35 has been replaced by the two-stage opamp of Figure 4.31. The opamp has been split into the first and the second stage: the first stage is linear, and its gain is  $A_{v1}(s)$ , the

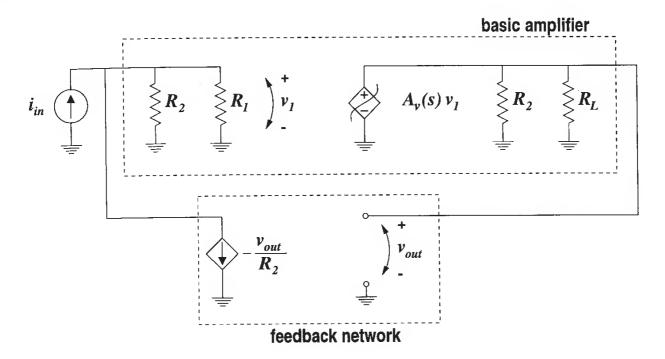


Figure 4.35: The inverting amplifier of Figure 4.32 split into a basic amplifier and a feedback network.

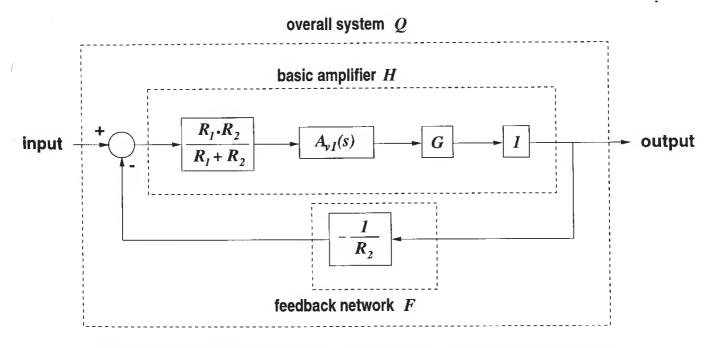


Figure 4.36: Final block diagram representation of the inverting amplifier.

value of which is given in equation (4.122). The second stage, which is nonlinear, is represented by the operator G.

The operation of the block diagram of Figure 4.36 can be interpreted as follows: the input, which is a current, is transformed into a voltage by the first block. This voltage corresponds to

the voltage  $v_1$  in Figure 4.35. This voltage is linearly amplified by the first stage of the opamp. The second stage of the opamp, represented by the operator G is nonlinear. Next, the output voltage is fed into the feedback network which transforms the voltage into a current.

The kernel transforms that correspond to G can be found from Figure 4.31 as follows: the nonlinear drain current of  $M_1$  is modeled with  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$ . The drain current is further transformed into a voltage by the internal load resistor  $R_{L_{int}}$ . Hence we find for  $G_1(s_1)$ ,  $G_2(s_1, s_2)$  and  $G_3(s_1, s_2, s_3)$ :

$$G_1(s_1) = G_1 = -g_m R_{L_{int}} (4.133)$$

$$G_2(s_1, s_2) = G_2 = -K_{2q_m} R_{L_{int}}$$
(4.134)

$$G_3(s_1, s_2, s_3) = G_3 = -K_{3q_m} R_{L_{int}}$$
(4.135)

Since capacitive effects in the second stage have been neglected, the arguments  $s_1$ ,  $s_2$  and  $s_3$  have been omitted. The factor  $G_1$  is dimensionless whereas  $G_2$  and  $G_3$  have a dimension of  $V^{-1}$  and  $V^{-2}$ , respectively.

We now have enough data to determine the kernel transforms of the complete basic amplifier of Figure 4.36, which is represented by the operator H. The transforms  $H_1(s_1)$ ,  $H_2(s_1, s_2)$  and  $H_3(s_1, s_2, s_3)$  of the basic amplifier are determined by applying the expressions for kernel transforms of a cascade connection, equations (4.68) through (4.70). To this purpose, the two linear blocks with transfer function  $(R_1R_2/(R_1 + R_2))$  and  $A_{v1}(s)$  are merged into one linear block. Then we find

$$H_1(s_1) = G_1 A_{v1}(s_1) \frac{R_1 R_2}{R_1 + R_2} = -g_m R_{L_{int}} A_{v1}(s_1) \frac{R_1 R_2}{R_1 + R_2}$$
(4.136)

$$H_2(s_1, s_2) = G_2 \left(\frac{R_1 R_2}{R_1 + R_2}\right)^2 A_{v1}(s_1) A_{v1}(s_2)$$

$$= -K_{2g_m} R_{L_{int}} \left(\frac{R_1 R_2}{R_1 + R_2}\right)^2 A_{v1}(s_1) A_{v1}(s_2)$$
(4.137)

$$H_{3}(s_{1}, s_{2}, s_{3}) = G_{3} \left(\frac{R_{1}R_{2}}{R_{1} + R_{2}}\right)^{3} A_{v1}(s_{1}) A_{v1}(s_{2}) A_{v1}(s_{3})$$

$$= -K_{3g_{m}} R_{L_{int}} \left(\frac{R_{1}R_{2}}{R_{1} + R_{2}}\right)^{3} A_{v1}(s_{1}) A_{v1}(s_{2}) A_{v1}(s_{3})$$

$$(4.138)$$

The dimensions of  $H_1(s_1)$ ,  $H_2(s_1, s_2, s_3)$  and  $H_3(s_1, s_2, s_3)$  are  $\Omega$ ,  $\Omega/A$  and  $\Omega/A^2$ , respectively. The kernel transforms  $F_1(s_1)$ ,  $F_2(s_1, s_2)$  and  $F_3(s_1, s_2, s_3)$  of the feedback network of Figure 4.36 are easily determined. One finds

$$F_1(s_1) = F_1 = -\frac{1}{R_2} \tag{4.139}$$

and since the feedback network is linear

$$F_2(s_1, s_2) = 0 (4.140)$$

$$F_3(s_1, s_2, s_3) = 0 (4.141)$$

Note that the dimension of  $F_1$  is  $\Omega^{-1}$ .

The gain reduction factor  $R(s_1)$  is found using equation (4.86):

$$R(s_1) = \frac{1}{1 + H_1(s_1)F_1(s_1)} \tag{4.142}$$

which, for a large loop gain  $H_1(s_1)F_1(s_1)$  reduces to

$$R(s_1) \approx \frac{1}{H_1(s_1)F_1(s_1)}$$
 (4.143)

Using equations (4.136) and (4.139) we obtain

$$R(s_1) \approx \frac{R_1 + R_2}{R_1 g_m R_{Lint} A_{v1}(s_1)} \tag{4.144}$$

It is seen that the gain reduction factor increases as the frequency increases. In other words, the gain is suppressed less as the frequency increases. However, since the gain of the opamp itself decreases as well, the closed-loop gain remains constant, at least until the gain bandwidth of the opamp.

Volterra kernels of the complete circuit and the harmonic distortion At this moment we have all parameters to determine the kernel transforms of the overall system  $\mathbf{Q}$ . Since we are only interested in determining the harmonic distortion figures of the amplifier, we know from equations (4.33) and (4.34) that we only need to know the kernel transforms for all frequency arguments being identical. In other words, we can limit ourselves to the determination of  $Q_2(s_1, s_1)$  and  $Q_3(s_1, s_1, s_1)$  rather than  $Q_2(s_1, s_2)$  and  $Q_3(s_1, s_2, s_3)$ . Using equations (4.93), (4.136) and (4.144) we find for the first-order transfer function of the overall feedback system

$$Q_1(s_1) = H_1(s_1)R(s_1) = -R_2 (4.145)$$

The dimension of  $Q_1(s_1)$  is  $\Omega$ . The second-order kernel transform is found using equations (4.94), (4.137) and (4.144):

$$Q_2(s_1, s_1) = (R(s_1))^2 R(2s_1) H_2(s_1, s_1)$$
(4.146)

$$=\frac{R_1+R_2}{R_1g_m^3R_{Lint}^3}\cdot\frac{R_2^2G_2}{A_{v1}(2s_1)}\tag{4.147}$$

The dimension of  $Q_2(s_1, s_1)$  is  $\Omega/A$ .  $Q_2(s_1, s_1)$  increases proportionally with the frequency. This can be seen at the occurrence of  $A_{v1}(2s_1)$  in the denominator: at frequencies sufficiently above the dominant pole of the opamp the gain of the first stage decreases proportionally to the frequency, as can be seen from equation (4.122). The linear increase of  $Q_2(s_1, s_1)$  with the frequency can be explained by the fact that the gain reduction factor  $R(s_1)$  increases linearly with the frequency whereas  $H_2(s_1, s_1)$  decreases with the square of the frequency.

Finally, the third-order kernel transform  $Q_3(s_1, s_1, s_1)$  is found using equations (4.95), (4.137). (4.138) and (4.144):

$$Q_{3}(s_{1}, s_{1}, s_{1}) = (R(s_{1}))^{3} \cdot \left[ H_{3}(s_{1}, s_{1}, s_{1}) - \frac{2H_{2}(s_{1}, s_{1})H_{2}(s_{1}, 2s_{1})}{H_{1}(2s_{1})} \right] \cdot R(3s_{1})$$

$$= \frac{R_{1} + R_{2}}{R_{1} g^{4} R_{2}^{4}} \cdot \frac{R_{2}^{3}}{A_{11}(3s_{1})} \left[ G_{3} - \frac{2G_{2}^{2}}{G_{1}} \right]$$

$$(4.149)$$

The dimension of  $Q_3(s_1, s_1, s_1)$  is  $\Omega/A^2$ . Just as for  $Q_2(s_1, s_1, s_1)$  it is seen that  $Q_3(s_1, s_1, s_1)$  increases linearly with the frequency.

The second and third harmonic distortion can now be determined using equations (4.33) and (4.34). To this purpose it is assumed that the input current source is a sinusoidal excitation of the form

$$i_{in}(t) = I_{in}\sin(\omega_1 t) \tag{4.150}$$

We recall that the actual input signal was not the current source  $i_{in}(t)$  but the voltage source  $v_{in}(t) = R_1 i_{in}(t)$ :

$$v_{in}(t) = V_{in}\sin(\omega_1 t) \tag{4.151}$$

and the relationship between the amplitudes  $V_{in}$  and  $I_{in}$  is given by

$$V_{in} = R_1 I_{in} (4.152)$$

For the second harmonic distortion we find

$$HD_2 = \frac{1}{2} \cdot I_{in} \cdot \left| \frac{Q_2(j\omega_1, j\omega_1)}{Q_1(j\omega_1)} \right| \tag{4.153}$$

$$= \frac{1}{2} \cdot \frac{V_{in}}{R_1} \cdot \frac{R_1 + R_2}{R_1 g_m^3 R_{L_{int}}^3} \cdot \frac{R_2 K_{2g_m} R_{L_{int}}}{|A_{v1}(2j\omega_1)|}$$
(4.154)

The factor  $|A_{v1}(2j\omega_1)|$  can be rewritten as

$$|A_{v1}(2j\omega_1)| = \left| \frac{A_{v10}p_d}{2j\omega_1} \right| = \frac{1}{g_m R_{L_{int}}} \frac{\pi GBW}{\omega_1}$$
 (4.155)

Then after some algebra we obtain

$$HD_2 = \frac{1}{2} \cdot V_{in} \cdot \frac{R_2}{R_1} \cdot \frac{R_1 + R_2}{R_1} \cdot K'_{2g_m} \cdot \frac{\omega_1}{g_m R_{L_{int}} \pi GBW}$$
(4.156)

It is seen that the second harmonic distortion increases proportionally to the frequency. This is due to the fact that  $Q_2(s_1, s_1)$  increases proportionally with the frequency, whereas  $Q_1(s_1)$ , which is nothing else but the closed-loop transresistance, remains constant.

Expression (4.156) is now evaluated for a practical example with  $R_1 = 1k\Omega$ ,  $R_2 = 20k\Omega$ . Then the low-frequency voltage gain of the complete amplifier is -20. Further, it is assumed that the gain-bandwidth product of the operational amplifier is 1MHz and the signal frequency is 10kHz. The gain of the second stage of the opamp,  $g_mR_{L_{int}}$ , is assumed to be 50. Transistor  $M_1$  is assumed to obey the simple square-law characteristic of a MOS transistor in saturation. Then from Table 3.2 we find the normalized nonlinearity coefficient  $K'_{2q_m}$ :

$$K_{2g_m}' = \frac{K_{2g_m}}{g_m} = \frac{1}{2(V_{GS} - V_T)} \tag{4.157}$$

Assuming a value of 0.2V for  $V_{GS}-V_T$  we find  $K'_{2g_m}=2.5V^{-1}$ . The input amplitude is assumed to be 100mV. This means that the output amplitude is about 2V. Using equation (4.156) the second harmonic distortion is then found to be 0.021 or 2.1%.

Next we consider the third harmonic distortion. Using equation (4.34) we find

$$HD_{3} = \frac{1}{4} \cdot I_{in} \cdot \left| \frac{Q_{3}(j\omega_{1}, j\omega_{1}, j\omega_{1})}{Q_{1}(j\omega_{1})} \right|$$
(4.158)

Using equations (4.145) and (4.149) and performing the same substitutions as for the second harmonic distortion we find

$$HD_{3} = \frac{1}{4} \cdot V_{in}^{2} \cdot \frac{R_{2}^{2}}{R_{1}^{1}} \cdot \frac{R_{1} + R_{2}}{R_{1}} \cdot \left| -K_{3g_{m}}' + 2K_{2g_{m}}'^{2} \right| \cdot \frac{3\omega_{1}}{2\pi GBW g_{m}^{2} R_{L_{int}}^{2}}$$
(4.159)

Just as with  $HD_2$ , it is seen that  $HD_3$  increases proportionally to the frequency. Next it is seen that the second-order nonlinearity of transistor  $M_1$  and the third-order nonlinearity partially cancel. This has also been seen in Section 4.8.3.2 with the analysis of the single-transistor amplifier with emitter degeneration.

An evaluation of equation (4.159) for the same numerical values that have been used for the evaluation of  $HD_2$ , yields  $HD_3 = 2.5 \times 10^{-5}$ . Hereby it has been assumed that transistor  $M_1$  is a perfect square-law device such that  $K_{3g_m}^t = 0$ . This is of course only an approximation. More accurate MOS models will be considered in Chapter 7.

#### 4.8.7 Nonlinear feedback applications

In Section 4.8.3, it was shown that linear feedback can reduce the distortion but it cannot completely linearize a given basic amplifier. This goal can be met, however, with nonlinear feedback in two ways.

A first means to linearization is by setting the expressions of  $Q_2(s_1, s_2)$  and  $Q_3(s_1, s_2, s_3)$  to zero, which provides conditions for the kernels  $\mathbf{F}_1$ ,  $\mathbf{F}_2$  and  $\mathbf{F}_3$  of the feedback network.

Under the assumption that the inverse  $\mathbf{H}_1^{-1}$  of  $\mathbf{H}_1$  exists and is stable, the second-order kernel  $\mathbf{Q}_2$  becomes zero if the second-order Volterra operator  $\mathbf{F}_2$  of the feedback network satisfies the relationship [Sche 80]

$$\mathbf{F}_2 = \mathbf{H}_1^{-1} \mathbf{H}_2 \mathbf{H}_1^{-1} \tag{4.160}$$

With this  $\mathbf{F}_2$ ,  $\mathbf{Q}_3$  can be made zero for  $\mathbf{F}_3$  given by

$$\mathbf{F}_{3} = \mathbf{H}_{1}^{-1}\mathbf{H}_{3}\mathbf{H}_{1}^{-1} - \mathbf{H}_{1}^{-1}\mathbf{H}_{2}\mathbf{H}_{1}^{-1}\left[\mathbf{H}_{1} + \mathbf{H}_{2}\right]\mathbf{H}_{1}^{-1} + \mathbf{H}_{1}^{-1}\mathbf{H}_{2}\mathbf{H}_{1}^{-1} + \mathbf{H}_{1}^{-1}\mathbf{H}_{2}\mathbf{H}_{1}^{-1}\mathbf{H}_{2}\mathbf{H}_{1}^{-1}$$
(4.161)

These results are not useful in practice for the linearization of a circuit, since the relationship between the operators of F and H is too complicated.

A more interesting way of linearization is shown in Figure 4.37. If the loop gain of the feedback loop in this schematic is sufficiently high and the nonlinearity of  $\mathbf{H}$  is negligible, then the input-output relationship is purely linear. The reason is that the feedback system mimics the inverse of the nonlinear feedback network  $\mathbf{F}$ , yielding exactly the required post-distortion to compensate for the nonlinearities of  $\mathbf{F}$ .

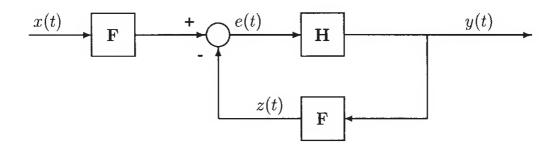


Figure 4.37: Cascade of a nonlinear network  $\mathbf{F}$  and a nonlinear feedback system with the same nonlinear network as a feedback network.

It is important to note that the principle schematic of Figure 4.37 can in practice only realize a linear voltage-to-voltage or a current-to-current conversion. If, for example, **F** is a (trans)conductance, then the feedback system is a (trans)resistance and vice versa.

The above principle is in fact the generalization of the active feedback described in [Gilb 74]: feedback with a large loop gain is used to compensate the nonlinearities of the voltage-to-current converters at the input of a Gilbert multiplier with pre-distortion.

#### 4.9 Summary

In this chapter weakly nonlinear analog integrated circuits have been analyzed conceptually. Volterra series are extremely useful to this purpose. Responses like harmonics and intermodulation products can be expressed in terms of Volterra kernels. Moreover, Volterra series allow to obtain insight into the system operation using a block diagram representation of the Volterra kernels. In this way, the following concepts have been studied that can be useful for analog integrated circuit design: suppression of even- or odd-order kernels, cascade connections of nonlinear systems, inverse nonlinear systems and, finally, linear and nonlinear feedback. From these analyses some design rules can be derived. First, it is seen that symmetry can be used to suppress either all even-order or all odd-order responses. The latter is not often used, since the

first-order response, which is usually the wanted signal, is suppressed as well. However, it is useful in mixers. Next it is seen that the output stages in amplifiers should be as linear as possible, since in general they give the largest contributions to the overall nonlinear distortion. Preor post-distortion is a means to suppress or reduce the distortion of a nonlinear circuit. In this way, a linear current-to-current or a voltage-to-voltage conversion can be achieved.

In most analog integrated circuits feedback is present, sometimes unwanted. The effect of feedback on second- and third-order responses is not very obvious since the responses of different order can interact in the feedback loop. These effects have been studied in detail with the model of a nonlinear basic amplifier which is fed back with a nonlinear feedback circuit. It is seen that the nonlinear distortion produced by the basic amplifier can be suppressed by a large loop gain if the feedback circuit is linear. Nonlinearities of the feedback circuit are not suppressed by the loop gain. Further, it is seen that in the presence of a large loop gain, a closed-loop system mimics the inverse of the feedback circuit, at least if the nonlinearities of the basic amplifier can be neglected. This can be exploited in pre- or post-distortion applications.

## **Chapter 5**

# Calculation of harmonics and intermodulation products

#### 5.1 Introduction

In this chapter different methods are explained for the calculation of harmonics and intermodulation products in a weakly nonlinear analog circuit.

Calculation methods for harmonics and intermodulation products have been reported several times either based upon Volterra series [Buss 74, Chua 79b, Wein 80, Lar 93, Maas 88]. As shown in Table 4.1, harmonics or intermodulation products can be obtained from the knowledge of the Fourier transform of Volterra kernels.

The different reported methods basically make use of the following approach for the calculation of Volterra kernels of node voltages and branch currents in a circuit. The Volterra kernels are computed by each time solving the same linear network that is nothing else but the linearized equivalent of the original circuit. This linear network is solved first to compute the first-order kernels. The second-order kernels are computed by exciting this linear network with inputs that are different from the external excitations. The new inputs depend on the first-order Volterra kernels. The third-order kernels are computed similarly, but again with different inputs. These inputs are now function of the first- and second-order kernels. Higher-order responses are computed similarly. For this approach the nonlinearities need to be described as basic nonlinearities, that have been discussed in Chapter 3. The approach computes every Volterra kernel as a sum of contributions, one for each basic nonlinearity. This approach is further explained in Section 5.2.

The computation of Volterra kernels from which harmonics and intermodulation products are derived, contains redundancy. Indeed, if, for example, one is interested in a third harmonic, then from Table 4.1 it is seen that this requires the knowledge of  $H_3(j\omega_1, j\omega_1, j\omega_1, j\omega_1)$ . Clearly, one is interested in the value of the third-order kernel evaluated for three identical arguments only, and not in a knowledge of the kernel  $H_3(j\omega_1, j\omega_2, j\omega_3)$  in which all three arguments are different. If, on the other hand, one is interested in inputs other than one or two sinusoidal signals then a complete knowledge of the third-order kernel might be desirable.

Not all circuits have only one input port and one output port. For example, many mixers have

two input ports. For such circuits, which in general are denoted as *multiple-input systems*, the use of Volterra series becomes awkward. For instance, the second-order kernel of a voltage is a matrix of size (# inputs × # inputs), the third-order kernel is a (# inputs × # inputs × # inputs) tensor, and so on [Sale 82, Sche 80, Chua 79b]. Calculations with tensors are quite cumbersome. Although it is possible to get rid of tensor manipulations by the use of the Kronecker product [Sale 82], an extension of the calculation method of [Buss 74, Chua 79b] in terms of Volterra series for multiple-input systems is complex.

In Section 5.3 a method is explained that directly computes the required response, which is a harmonic or an intermodulation product. Both the redundancy of the Volterra series approach and the use of tensors for multiple-input systems are avoided. The method circumvents the use of Volterra series. Of course, the resulting harmonics or intermodulation products obtained with this method are identical to the results obtained with the Volterra series approach. The method also computes the responses by repeatedly solving a linearized network, just as with the Volterra series approach. Details of the derivation of the method are given in Appendix C. In Section 5.3 this method is explained with an example. It will be seen again that the response of interest is a sum of contributions, one for each basic nonlinearity.

A third computation method that again leads to the same result has been described in [Kuo 73, Chis 73, Chua 75]. This method can be considered as a perturbation method. Since this method leads to the same results, it is not considered here.

Since with the different methods mentioned above, a harmonic or intermodulation product can be considered as a sum of contributions, the basic nonlinearities that give the dominant contributions to a harmonic or an intermodulation product can be distinguished. The extraction of the dominant contribution yields valuable information. First, it indicates which basic nonlinearities are mainly responsible for the observed nonlinear behavior at the output of the circuit. Next, the dominant contributions can be analyzed in the frequency domain, in much the same way as an AC analysis of a linearized circuit. The results of such analysis, as performed by numerous commercial circuit simulators, are usually presented in the form of a plot, that shows the response of interest as a function of frequency. We will use the same approach in Chapter 8.

However, the insight in the nonlinear operation of a circuit would be greatly enhanced if closed-form, symbolic expressions could be generated for the contributions of the basic nonlinearities that are dominant in the harmonic or intermodulation product. Such expressions can, at least theoretically, be obtained by hand calculations since they can be obtained by solving linear networks. However, for circuits of practical size it will become clear in this chapter that hand calculations become very complicated. Here *symbolic network analysis programs* can be very useful to speed up or to partially automate the calculations. Such programs can generate closed-form expressions for the AC characteristics of an analog circuit [Giel 91, Giel 94a]. In order to generate compact, interpretable expressions, the symbolic network analysis programs can generate approximate expressions by removing the small terms of an expression according to a user-defined accuracy.

Symbolic network analysis programs have been developed originally to compute expressions of linearized circuits. Since the calculation of Volterra kernels or harmonics and intermodulation products reduces to a repeated solution of sets of linear equations, the kernel of a symbolic network analysis program can be used to obtain closed-form expressions for the nonlinear behavior

of a circuit. This issue is explained in Section 5.4. The symbolic analysis program that is used in this book to compute nonlinear responses is the program ISAAC [Giel 89, Giel 91].

The calculation methods explained in this Chapter are illustrated with several examples. The Volterra series approach that is explained in Section 5.2 is illustrated with computations on a common emitter amplifier. The direct calculation of nonlinear responses, explained in Section 5.3, is illustrated with a simple mixer, both in bipolar and in CMOS. Next, in Section 5.5 two simple circuits are analyzed, namely a resistive voltage divider and a capacitive current divider. The results obtained with these circuits yield insights that are useful for the analyses in Chapter 8.

As already mentioned above, the Volterra series approach can only compute a reasonable approximation of a circuit's nonlinear behavior when the circuit behaves in a weakly nonlinear way. In practice it is not always easy to predict the limit of the signal amplitudes to which the weakly nonlinear approximation is valid. In order to check the validity of the assumption that the circuit behaves in a weakly nonlinear way, iterative numerical simulation techniques must be used. Some of these techniques are discussed in Section 5.6.

Before a comprehensive discussion of the Volterra series approach and the direct calculation of nonlinear responses, it is already possible to get a feeling of how these methods proceed by considering the example of how a second-order nonlinearity in a circuit contributes to the second harmonic at the output. In Chapter 4 it has been explained that a second-order nonlinearity combines two signals, which eventually are identical, and produces a signal the order of which is the sum of the order of the two signals. Further, we know that the second harmonic at the output of a weakly nonlinear circuit is primarily determined by second-order behavior. Such behavior can only be caused by second-order nonlinearities that combine two first-order signals. Hence, the contribution of a second-order nonlinearity to the overall second harmonic can be computed as follows: the first-order response of the voltage or current that controls the nonlinearity is computed first. Higher-order responses do not have to be computed since they will result in a signal, of order higher than two when they are fed into the second-order nonlinearity and combined with another signal. This first-order response is simply computed using linear network analysis. This response is then squared by the second-order nonlinearity — which in this case is equivalent to saying that the nonlinearity combines two identical signals to produce a second-order signal and the resulting signal propagates further to the output of the circuit. Hereby the influence of other nonlinearities in the path from the second-order nonlinearity to the output, should not be considered. Indeed, these would combine the second-order signal with other signals to produce a signal of order higher than two. In other words, one must compute the propagation of the second-order signal through the linearized network, which again can be performed with linear network analysis. The computation method is illustrated schematically in Figure 5.1.

For a third-order nonlinearity a similar reasoning can be followed to find the contribution of this nonlinearity to the third harmonic at the output. The difference with a second-order nonlinearity is that a third-order nonlinearity combines three signals instead of two to produce a signal the order of which is the sum of the orders of the three individual signals.

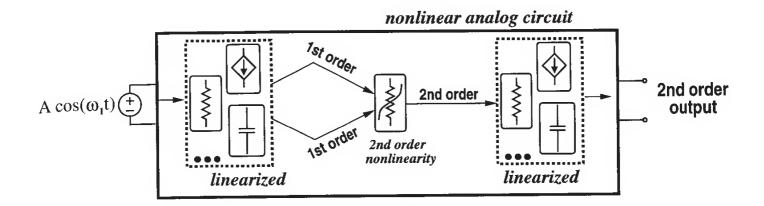


Figure 5.1: Illustration of the calculation method for second-order kernel transforms or harmonics or intermodulation products. For the computation of the second harmonic, the two first-order signals are identical.

Prerequisites for the calculation methods For both the Volterra series approach and the direct computation of harmonics and intermodulation products it is assumed that the power series description of the different basic nonlinearities that are present in a circuit, can be broken down after the first few terms without a significant loss of accuracy. Further it is necessary that every voltage and current in the circuit can be described by a converging Volterra series. Finally, the circuit must have a unique solution for every node or branch voltage and branch current.

The calculation method is explained for circuits with voltage-controlled nonlinearities. Current-controlled nonlinearities such as nonlinear resistors can most often be described as nonlinear conductances as explained in Section 3.2.2.1. Other current-controlled nonlinearities are seldom used in analog integrated circuit design. Also, nonlinear voltage-controlled voltage sources are not considered in this chapter. Nevertheless, an extension of the calculation methods explained in this chapter to all these elements does not yield any difficulties.

## 5.2 Calculation of Volterra kernels

In this section it is explained how Volterra kernels of order one, two and three of voltages and currents in a weakly nonlinear circuit can be computed. The method is explained using an example circuit. Detailed derivations can be found in the literature [Buss 74, Chua 79b].

The method computes the Volterra kernels or the responses in increasing order by repeatedly solving a linear network. First, the linearized circuit is analyzed with the external excitation(s) applied. For higher orders the same linearized circuit is solved with other inputs.

The example circuit is shown in Figure 5.2. This is an equivalent scheme for a single-transistor (BJT) amplifier loaded with a resistance and a capacitance and excited by a voltage source. The base current is the current  $i_1$ . The collector current is the current  $i_2$ . The nonlinearity of these two currents will be taken into account in the calculations. The excitation is a voltage source  $v_{in}(t)$ . For the determination of the Volterra kernels the actual shape of the input signal in fact is not important, as long as the input amplitude remains small enough such that the circuit

behaves in a weakly nonlinear way.

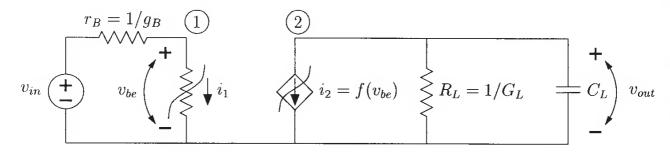


Figure 5.2: Equivalent circuit of a bipolar transistor with a load resistance and a load capacitance at its collector.

### 5.2.1 First-order kernels

In a first step, the response of the linearized circuit to the external inputs is calculated as a function of frequency. This yields the first-order Volterra kernels, which are nothing else but transfer functions of the linearized circuit. For this computation every nonlinearity is replaced with its linearized equivalent. Together with the output voltage, all voltages that control a nonlinearity are calculated as well. The reason for these extra computations will become clear when the calculation of higher-order responses is discussed.

The calculation of the first-order responses for the example of Figure 5.2 is illustrated in Figure 5.3.

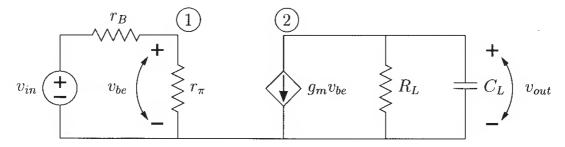


Figure 5.3: Linearized equivalent of the circuit of Figure 5.2.

In general this calculation can be represented by the solution of the following matrix equation:

$$\mathbf{Y}(s).\mathbf{H}_1(s) = \mathbf{I}\mathbf{N}_1 \tag{5.1}$$

in which Y(s) is the admittance matrix of the circuit,  $H_1(s)$  is the vector of first-order Volterrakernel transforms of the node voltages and  $IN_1$  is the vector of excitations. The admittance matrix results from the application of Kirchoff's current law at every node voltage. This way of writing down the network equations is denoted as **nodal analysis** [Chua 75]. The unknowns that result from this formulation are the node voltages. If a network contains zero-impedance

elements such as voltage sources, then an extra unknown needs to be introduced, namely the current through this zero-impedance element. Also, an extra equation is provided for the branch with the zero-impedance element. The resulting formulation is called *modified nodal analysis* (MNA). This type of equation formulation is explained in many textbooks on network analysis [Chua 75, Chua 87]. For the MNA formulation, Y(s) is the MNA matrix and  $H_1(s)$  is the vector of the first-order kernels of every node voltage and of some branch currents. The circuit output and the voltages that control a nonlinearity can in general be written as a linear combination of the elements of vector  $H_1(s)$ .

In many cases the unknown currents that result from the presence of zero-impedance elements can be eliminated in advance. Then the method for the formulation of equations is referred to as the compacted MNA method (CMNA) [Giel 89, Giel 91]. This formulation method will be illustrated with the example circuit of Figure 5.3.

Applying Kirchoff's current law at node 1 in Figure 5.3 yields:

$$g_B(V_1 - V_{in}) + g_\pi V_1 = 0 (5.2)$$

At node 2 we obtain in the same way

$$g_m V_{be} + (G_L + sC_L)V_2 = 0 (5.3)$$

The voltages in equations (5.2) and (5.3) are Laplace transforms. If we would have taken the MNA method instead of the CMNA method to formulate the network equations, then we would have to write down an extra equation at the positive node of the voltage source. This equation would then include the current through the voltage source. In addition, a branch relation would be required stating that the voltage at the positive node of the voltage source is equal to  $V_{in}$ . Clearly, the CMNA method yields less network equations than the MNA method.

Equations (5.2) and (5.3) can be combined in one matrix equation:

$$\begin{bmatrix} g_B + g_{\pi} & 0 \\ g_m & g_L + sC_L \end{bmatrix} \cdot \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} g_B V_{in} \\ 0 \end{bmatrix}$$
 (5.4)

The  $2\times 2$ -matrix in the left-hand side is the (C)MNA matrix. It is clear that  $V_1$  and  $V_2$  reduce to the transfer functions of the voltages at node 1 and 2 when  $V_{in}$  is set equal to one. These transfer functions are denoted by  $H_{1_1}(s)$  and  $H_{1_2}(s)$ . The first subscript in these two transfer functions indicates the order of the transfer function, whereas the second subscript corresponds to the numbering of the node voltages. Hence these transfer functions are found from the matrix equation

$$\begin{bmatrix} g_B + g_{\pi} & 0 \\ g_m & g_L + sC_L \end{bmatrix} \cdot \begin{bmatrix} H_{1_1}(s) \\ H_{1_2}(s) \end{bmatrix} = \begin{bmatrix} g_B \\ 0 \end{bmatrix}$$
 (5.5)

This set of equations can be solved using Cramer's rule. In this way,  $H_{1_1}(s)$  and  $H_{1_2}(s)$  are found as a ratio of two determinants, the denominator of both being the determinant of the CMNA matrix, denoted by  $\det(s)$ 

$$\det(s) = (g_L + sC_L)(g_B + g_{\pi})$$
(5.6)

Then  $H_{1_1}(s)$  is given by

$$H_{1_1}(s) = \frac{\begin{vmatrix} g_B & 0 \\ 0 & g_L + sC_L \end{vmatrix}}{\det(s)} = \frac{g_B(g_L + sC_L)}{\det(s)} = \frac{g_B}{g_B + g_\pi}$$
 (5.7)

This transfer function could have been written down immediately since the voltage at node 1 is just a fraction  $r_{\pi}/(r_{\pi}+r_B)$  of the input voltage. Equation (5.7) is the same but in terms of conductances instead of resistances.

The linear transfer function at node 2 is given by

$$H_{1_2}(s) = \frac{\begin{vmatrix} g_B + g_\pi & g_B \\ g_m & 0 \end{vmatrix}}{\det(s)} = \frac{-g_m g_B}{\det(s)} = \frac{-g_m g_B}{(g_B + g_\pi) (g_L + sC_L)}$$
(5.8)

Again, this transfer function could have been written down without making the above calculations, but we have followed here a systematic way of deriving  $H_{1_1}(s)$  and  $H_{1_2}(s)$  in order to better see the parallel with the computation of higher-order responses as will be done below.

### 5.2.2 Second-order kernels

Having computed the first-order kernels, which are nothing else but linear transfer functions, the second-order Volterra kernels can be calculated with the method that is explained in [Buss 74]. To this purpose, the excitation is first put to zero in the linearized circuit. This means that  $v_{in}$  is replaced by a short circuit. The second-order kernel transforms of the node voltages and branch currents in the nonlinear circuit are found by solving the same linearized network as the one that was used for the calculation of linear transfer functions but now with other inputs: instead of the real excitation, the so-called *nonlinear current sources of order two* or second-order nonlinear current sources are applied to the linearized circuit. The node voltages and branch currents that are found in this way are equal to the second-order kernel transforms.

Every nonlinearity in the original circuit gives rise to such a current source in the linearized circuit. The sources are placed in parallel with each nonlinear element that has been linearized. The orientation of the sources is the same as the orientation of the controlled current in the original nonlinear circuit. In fact, the nonlinear current sources model the corrections on the linear response for second-order nonlinearities.

The computation of the second-order kernels using the second-order nonlinear current sources is shown for the example circuit in Figure 5.4: the original nonlinear circuit of Figure 5.2 has two nonlinearities, namely a nonlinear conductance that represents the nonlinear base current and a transconductance corresponding to the nonlinear collector current. These nonlinearities each give rise to a nonlinear current source.

The value of the current sources depends on the type of the basic nonlinearity, on its second-order nonlinearity coefficient and on the first-order kernels of the controlling voltage(s). The dependence on these first-order kernels explains why the first-order kernels have to be computed prior to the second-order kernels. The expressions for the nonlinear current sources are given

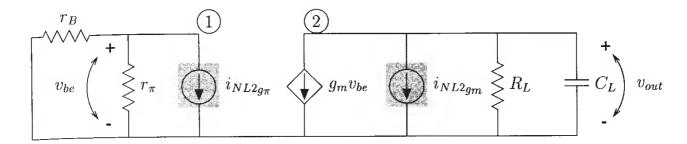


Figure 5.4: Circuit that has to be solved for the computation of second-order kernels in the circuit of Figure 5.2.

in Table 5.1. For both a nonlinear conductance and a nonlinear transconductance the value of the nonlinear current source is equal to the second-order nonlinearity coefficient, multiplied with the square of the first-order transfer function of the voltage that controls the nonlinearity. For a conductance this is the voltage over the element itself.

For a nonlinear capacitance the value of the nonlinear current source is similar as for a nonlinear conductance, except that it is multiplied with the frequency  $(s_1 + s_2)$ .

For a two-dimensional nonlinearity we recall that its power series can be split into two series that are identical to a one-dimensional conductance and a series that contains nothing but cross-terms. Table 5.1 shows only the component of the nonlinear current source that corresponds to the series with nothing but cross-terms. For a three-dimensional conductance only the terms are considered that correspond to the series that contains cross-terms in three voltages at the same time. The other components of the nonlinear current sources are similar to those of one- and two-dimensional conductances. Since a term that contains three different voltages at the same time is at least of order three, the second-order nonlinear current source that corresponds to such third-order nonlinearity is zero.

We only considered nonlinear voltage-controlled elements in admittance form in our discussions, which means that the current through a nonlinear element is a function of one or more voltages. If second-order kernel transforms would have to be computed in a circuit with nonlinear voltage-controlled voltage sources, then instead of applying nonlinear current sources of order two in parallel with the linearized equivalent of the nonlinearity, a nonlinear voltage source of order two would have to be applied in series with the controlled voltage source. This is explained in [Chua 79a, Chua 79b].

The computation of the second-order kernel transforms corresponds to the solution of the following matrix equation:

$$Y(s_1 + s_2).H_2(s_1, s_2) = IN_2$$
(5.9)

In this equation, the vector  $\mathbf{H}_2(s_1, s_2)$  represents the second-order kernel transforms of the node voltages and some currents in the circuit. Vector  $\mathbf{IN}_2$  corresponds to the nonlinear current sources of order two. Comparing equation (5.1) to equation (5.9), one can see that the same admittance or (C)MNA matrix is used, but it is evaluated at the frequency  $s_1 + s_2$  instead of s.

The computation method for the second-order kernels can be interpreted as follows: every second-order nonlinearity in the circuit combines two first-order components of its controlling

type of basic nonlinearity	expression for nonlinear current source of order two
(trans)conductance	$K_{2g_1}H_{1k}(s_1)H_{1k}(s_2)$
capacitance	$(s_1 + s_2) K_{2C_1} H_{1k}(s_1) H_{1k}(s_2)$
two-dimensional	$ \frac{1}{2} \left[ K_{2g_1 \& g_2} H_{1k}(s_1) H_{1l}(s_2) \right] $
conductance	, L
(only cross-terms)	$+K_{2g_1\&g_2}H_{1k}(s_2)H_{1l}(s_1)$
three-dimensional	
conductance	0
(only cross-terms)	

Table 5.1: Nonlinear second-order current sources for the different basic nonlinearities. These sources are used for the calculation of the second-order Volterra kernels.  $H_{1k}(\cdot)$  is the first-order transfer function of the voltage that controls a one-dimensional nonlinearity or the first controlling voltage of a two-dimensional conductance;  $H_{1l}(\cdot)$  is the first-order transfer function of the second voltage that controls the two-dimensional conductance.

voltage(s) to produce a second-order signal. This signal, at frequency  $s_1 + s_2$ , propagates through the rest of the linearized circuit to the output and to all other voltages and currents in the circuit yielding the demanded second-order kernels. The effect of nonlinearities does not have to be considered here since these would yield responses of order higher than two.

Once the second-order Volterra kernels are known, the second-order harmonics and intermodulation products can be computed using Table 4.1 which gives the relationship between the Volterra kernels and the second- and third-order responses.

We now apply the procedure to the computation of the second-order kernel transforms in the circuit of Figure 5.2. To this purpose, the circuit is linearized and the nonlinear current sources of order two are applied. The result is shown in Figure 5.4. Applying Kirchoff's current law at nodes 1 and 2 in Figure 5.4 leads to the matrix equation

$$\begin{bmatrix} g_B + g_{\pi} & 0 \\ g_m & g_L + (s_1 + s_2)C_L \end{bmatrix} \cdot \begin{bmatrix} H_{2_1}(s_1, s_2) \\ H_{2_2}(s_1, s_2) \end{bmatrix} = \begin{bmatrix} -i_{NL2g_{\pi}} \\ -i_{NL2g_m} \end{bmatrix}$$
 (5.10)

It is seen that the leftmost matrix is nothing else but the CMNA matrix of equation (5.5), but now evaluated at  $(s_1 + s_2)$  instead of at  $s_1$ . In the matrix equation (5.10) the unknowns are the second-order kernels of the voltages at nodes one and two,  $H_{2_1}(s_1, s_2)$  and  $H_{2_2}(s_1, s_2)$ , respectively. The

value of the nonlinear current source  $i_{NL2g_{\pi}}$  is obtained according to Table 5.1:

$$i_{NL2q_{\pi}} = K_{2q_{\pi}} H_{1_1}(s_1) H_{1_1}(s_2) \tag{5.11}$$

Using equation (5.7) we find

$$i_{NL2g_{\pi}} = \frac{K_{2g_{\pi}}g_B^2 (g_L + s_1 C_L) (g_L + s_2 C_L)}{\det(s_1) \det(s_2)}$$
(5.12)

Similarly, we find for  $i_{NL2gm}$ 

$$i_{NL2g_m} = K_{2g_m} H_{1_1}(s_1) H_{1_1}(s_2) = \frac{K_{2g_m} g_B^2 (g_L + s_1 C_L) (g_L + s_2 C_L)}{\det(s_1) \det(s_2)}$$
(5.13)

Note that the dimension of the nonlinear current sources in this example is  $A/V^2$ .

Using the values of the nonlinear current sources, the second-order transfer functions can be found by applying Cramer's rule on the matrix equation (5.10). In this way, we find for  $H_{2_1}(s_1, s_2)$ 

$$H_{2_1}(s_1, s_2) = \frac{\begin{vmatrix} -i_{NL2g_{\pi}} & 0\\ -i_{NL2g_{m}} & g_L + (s_1 + s_2)C_L \end{vmatrix}}{\det(s_1 + s_2)}$$
(5.14)

$$= \frac{-(g_L + (s_1 + s_2)C_L)}{\det(s_1 + s_2)} i_{NL2g_{\pi}}$$
 (5.15)

Using equation (5.12) this reduces to

$$H_{2_1}(s_1, s_2) = -\frac{(g_L + (s_1 + s_2)C_L)(g_L + s_1C_L)(g_L + s_2C_L)K_{2g_{\pi}}g_B^2}{\det(s_1 + s_2)\det(s_1)\det(s_2)}$$
(5.16)

This can be further simplified using the expression of the determinant of the (C)MNA matrix, equation (5.6):

$$H_{2_1}(s_1, s_2) = -\frac{K_{2g_{\pi}} g_B^2}{(g_B + g_{\pi})^3}$$
(5.17)

It is seen that the dimension of  $H_{2_1}(s_1, s_2)$  is  $V^{-1}$ .

For the second-order kernel transform at node 2 we proceed in the same way:

$$H_{22}(s_1, s_2) = \frac{\begin{vmatrix} g_B + g_\pi & -i_{NL2}g_\pi \\ g_m & -i_{NL2}g_m \end{vmatrix}}{\det(s_1 + s_2)}$$
(5.18)

$$= \frac{-(g_B + g_\pi) i_{NL2g_m} + g_m i_{NL2g_\pi}}{\det(s_1 + s_2)}$$
 (5.19)

Using equations (5.6), (5.12) and (5.13) this reduces to

$$H_{22}(s_1, s_2) = \frac{-g_B^2 \left( (g_B + g_\pi) K_{2g_m} - g_m K_{2g_\pi} \right)}{\left( g_B + g_\pi \right)^3 \left( g_L + (s_1 + s_2) C_L \right)}$$
(5.20)

The interpretation of the results is postponed until Section 5.2.4 and Chapter 8, where a single-transistor amplifier is analyzed in detail.

### 5.2.3 Third-order kernels

In the following step, the third-order transfer functions are calculated. Just as second-order ones, they are computed as the response to nonlinear current sources, this time of order three. Hence, a similar matrix equation must be solved as for order one and two:

$$\mathbf{Y}(s_1 + s_2 + s_3).\mathbf{H}_3(s_1, s_2, s_3) = \mathbf{IN}_3 \tag{5.21}$$

The expressions of the nonlinear current sources of order three are presented in Table 5.2. Clearly, the expressions are more involved than for the second order.

type of basic	expression for nonlinear current source	
nonlinearity	of order three	
(trans)conductance	$K_{3g_1}H_{1k}(s_1)H_{1k}(s_2)H_{1k}(s_3)$ $+\frac{2}{3}K_{2g_1}\left[H_{1k}(s_1)H_{2k}(s_2,s_3) + H_{1k}(s_2)H_{2k}(s_1,s_3) + H_{1k}(s_3)H_{2k}(s_1,s_2)\right]$	
capacitance	$(s_{1} + s_{2} + s_{3})K_{3}_{C_{1}}H_{1k}(s_{1})H_{1k}(s_{2})H_{1k}(s_{3})$ $+\frac{2}{3}(s_{1} + s_{2} + s_{3})K_{2}_{C_{1}}\left[H_{1k}(s_{1})H_{2k}(s_{2}, s_{3})$ $+H_{1k}(s_{2})H_{2k}(s_{1}, s_{3}) + H_{1k}(s_{3})H_{2k}(s_{1}, s_{2})\right]$	
two-dimensional	$ \frac{1}{3}K_{2g_{1}\&g_{2}}\left[H_{1k}(s_{1})H_{2l}(s_{2},s_{3}) + H_{1k}(s_{2})H_{2l}(s_{1},s_{3}) + H_{1k}(s_{3})H_{2l}(s_{1},s_{2}) + H_{2k}(s_{1},s_{2})H_{1l}(s_{3}) + H_{2k}(s_{1},s_{3})H_{1l}(s_{2}) + H_{2k}(s_{2},s_{3})H_{1l}(s_{1})\right] + \frac{1}{3}K_{3g_{1}\&g_{2}}\left[H_{1k}(s_{1})H_{1k}(s_{2})H_{1l}(s_{3}) + H_{1k}(s_{1})H_{1k}(s_{3})H_{1l}(s_{2})\right] $	
conductance (only cross-terms)	$egin{align*} &+ rac{1}{3} K_{32g_1\&g_2} \left[ H_{1k}(s_1) H_{1k}(s_2) H_{1l}(s_3) + H_{1k}(s_1) H_{1k}(s_3) H_{1l}(s_1)  ight] \\ &+ H_{1k}(s_2) H_{1k}(s_3) H_{1l}(s_1)  ight] \\ &+ rac{1}{3} K_{3g_1\&2g_2} \left[ H_{1k}(s_1) H_{1l}(s_2) H_{1l}(s_3) + H_{1k}(s_2) H_{1l}(s_3) + H_{1k}(s_2) H_{1l}(s_1) H_{1l}(s_2)  ight] \\ &+ H_{1k}(s_2) H_{1l}(s_1) H_{1l}(s_3) + H_{1k}(s_3) H_{1l}(s_1) H_{1l}(s_2)  ight] \end{aligned}$	

Table 5.2: Nonlinear third-order current sources for the different basic nonlinearities. These sources are used for the calculation of third-order Volterra kernels.  $H_{1k}(\cdot)$  and  $H_{2k}(\cdot)$  are the first- and second-order transfer functions of the voltage that controls a one-dimensional nonlinearity or of the first controlling voltage of a two-dimensional conductance;  $H_{1l}(\cdot)$  and  $H_{2l}(\cdot)$  are the first-order transfer functions of the second voltage that controls the two- and three-dimensional conductances;  $H_{1l}(\cdot)$  and  $H_{2l}(\cdot)$  are the first-order transfer functions of the third voltage that controls the three-dimensional conductance.

The nonlinear current sources for the one-dimensional nonlinearities consist of two components. The first component is caused by the third-order nonlinearity. This nonlinearity combines three first-order signals to a third-order one at frequency  $(s_1 + s_2 + s_3)$ . The second component is the average of terms which only differ by their arguments. This averaging is required to make the Volterra kernels symmetric with respect to their arguments. This component is caused by the second-order nonlinearity which acts upon a second-order signal and a first-order signal at the controlling voltage. Both third-order signals — with frequency  $(s_1 + s_2 + s_3)$  — propagate through the circuit to the output and to all other voltages and currents. During this propagation only the linear behavior of the circuit needs to be considered since interactions with other nonlinearities result into behavior of order higher than three.

For a two- and three-dimensional conductance, again the only component listed is the one which corresponds to the series of nothing but cross-terms with two and three voltages, respectively. The nonlinear current source that corresponds to the cross-terms of a two-dimensional nonlinearity consists of three components which are again averages of terms which only differ by their arguments. The first component is caused by the second-order nonlinearity  $K_{2g_1\&g_2}$  which acts upon a first-order signal of the first controlling voltage and a second-order signal of the second one and vice versa. The second component is caused by the third-order nonlinearity  $K_{3g_1\&g_2}$  which produces a third-order signal from two first-order signals of the first controlling voltage together with a first-order signal of the second controlling voltage. Coefficient  $K_{3g_1\&g_2}$  does exactly the same with the roles of the first and the second controlling voltage reversed.

The nonlinear current source associated with the cross-terms in three controlling voltages of a three-dimensional nonlinearity is caused by the third-order coefficient  $K_{3g_1\&g_2\&g_3}$ . It is again the average of terms in which the arguments are interchanged. This third-order nonlinearity takes together the first-order signals at the three controlling voltages to produce a third-order signal.

Once the third-order transfer functions are known, all third-order responses listed in Table 4.1 can be computed.

The application of the calculation method to the example of Figure 5.2 leads to the set of equations

$$\begin{bmatrix} g_B + g_{\pi} & 0 \\ g_m & g_L + (s_1 + s_2 + s_3)C_L \end{bmatrix} \cdot \begin{bmatrix} H_{3_1}(s_1, s_2, s_3) \\ H_{3_2}(s_1, s_2, s_3) \end{bmatrix} = \begin{bmatrix} -i_{NL3}g_{\pi} \\ -i_{NL3}g_m \end{bmatrix}$$
(5.22)

in which  $H_{3_1}(s_1, s_2, s_3)$  and  $H_{3_2}(s_1, s_2, s_3)$  are the third-order kernels of the voltages at node one and two, respectively. The above matrix equation results from the application of Kirchoff's current law at nodes 1 and 2, just as we did for the computation of the second-order kernel transforms. It is seen that the matrix equation (5.22) resembles very well to the matrix equation (5.10), except that now nonlinear current sources of order three are applied and the CMNA matrix is evaluated for  $(s_1 + s_2 + s_3)$  instead of  $(s_1 + s_2)$ .

From equation (5.22) we find, using the rule of Cramer

$$H_{3_1}(s_1, s_2, s_3) = \frac{-(g_L + (s_1 + s_2 + s_3)C_L)i_{NL3g_{\pi}}}{\det(s_1 + s_2 + s_3)}$$
(5.23)

$$H_{3_2}(s_1, s_2, s_3) = \frac{-(g_B + g_\pi) i_{NL3g_m} + g_m i_{NL3g_\pi}}{\det(s_1 + s_2 + s_3)}$$
(5.24)

A further elaboration of these expressions is again postponed until Section 5.2.4 and Chapter 8 where an in-depth analysis of a one-transistor amplifier is presented.

For higher-order transfer functions, a similar procedure can be followed. The expressions for the nonlinear current sources become more and more complicated and they depend on the lower-order kernels of the voltage(s) that control(s) the nonlinearity.

The complete calculation method is summarized in Figure 5.5. In this flowchart it is seen that for each order the same linearized network is solved. The node voltages that are found at each computation are equal to the Volterra kernels of the order under consideration. The wanted responses are derived from the Volterra kernels by using the relationship between the kernels and the nonlinear responses as given in Table 4.1. This will be illustrated in the next section.

### **5.2.4** Postprocessing of the results

The calculation method explained yields Volterra kernel transforms of node voltages or currents in a nonlinear network. In order to obtain the harmonic or intermodulation product of interest, the relationships between Volterra kernels and the nonlinear response of interest must be taken into account, as given in Table 4.1. The final results contain the small-signal parameters and the nonlinearity coefficients  $K_2$  and  $K_3$ , which are quantities that do not sound familiar yet to circuit engineers. However, if a designer wants to reason about nonlinear circuits he has to become acquainted with these parameters. Just as small-signal parameters, the nonlinearity coefficients can be expressed in terms of bias voltages or currents, model parameters and physical constants.

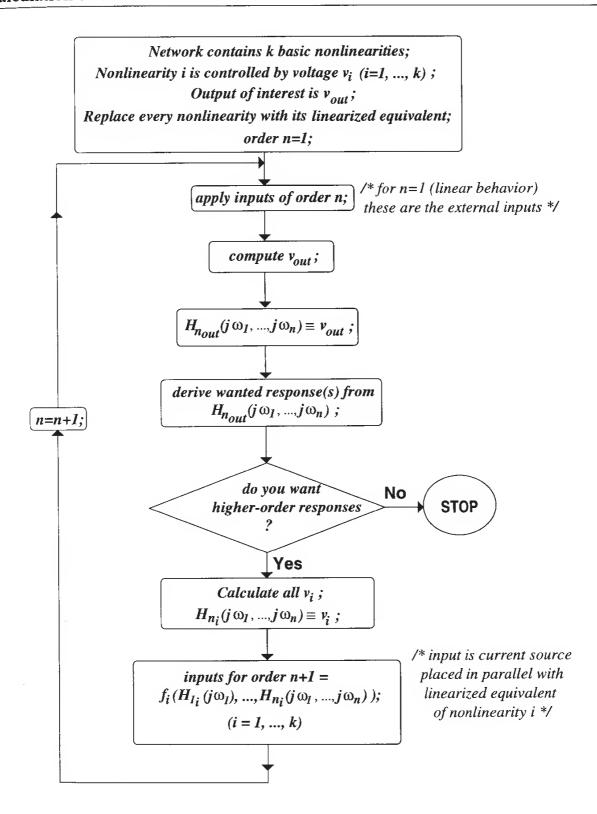


Figure 5.5: Schematic representation of the method of [Buss 74] for the computation of Volterra kernels.

Some simple examples of such expressions have already been presented in Chapter 3. More accurate values for the nonlinearity coefficients of the different nonlinearities in a bipolar and a

MOS transistor will be discussed in Chapters 6 and 7.

We now proceed to the postprocessing of the results of the example circuit of Figure 5.2 obtained in Sections 5.2.2 and 5.2.3. Assume that  $v_{in}$  in Figure 5.2 is a sinusoidal voltage of the form

$$v_{in} = V_{in}\sin(\omega_1 t) \tag{5.25}$$

The second harmonic has to be derived from the second-order kernel  $H_{2_2}(s_1, s_2)$  at node 2. From Table 4.1 it is seen that the complex amplitude of the second harmonic is found as

$$V_{2,2,0} = \frac{V_{in}^2}{2} H_{22}(j\omega_1, j\omega_1)$$
 (5.26)

The notation with the three subscripts must be interpreted as follows:  $V_{i,m,n}$  is the phasor of the component of the voltage on node i at the frequency  $m\omega_1 + n\omega_2$ .

Using equation (5.20) we find

$$V_{2,2,0} = \frac{V_{in}^2}{2} \cdot \frac{-g_B^2 \left( (g_B + g_\pi) K_{2g_m} - g_m K_{2g_\pi} \right)}{\left( g_B + g_\pi \right)^3 \left( g_L + 2j\omega_1 C_L \right)}$$
(5.27)

This expression can be further interpreted by substituting the nonlinearity coefficients by values that depend on bias currents, model parameters and technological constants. Using the simple values for  $K_{2g_m}$  and  $K_{2g_\pi}$  obtained in equations (3.14) and (3.22), equation (5.27) further reduces to

$$V_{2,2,0} = -\frac{g_B^3}{4V_t (g_B + g_\pi)^3} \frac{g_m}{(g_L + 2j\omega_1 C_L)} V_{in}^2$$
(5.28)

If the base resistance is zero, which means that  $g_B$  goes to infinity, then the complex amplitude of the second harmonic becomes

$$V_{2,2,0} = -\frac{1}{4V_t} \cdot \frac{g_m}{(g_L + 2j\omega_1 C_L)} V_{in}^2$$
 (5.29)

The second harmonic distortion is found by dividing  $V_{2,2,0}$  by the first-order response at node 2. Using equation (5.8) we find

$$HD_2 = \frac{V_{in}}{4V_t} \cdot \frac{g_L + j\omega_1 C_L}{g_L + 2j\omega_1 C_L} \tag{5.30}$$

At low frequencies, where the impedance of the load capacitance is much higher than the load resistance,  $HD_2$  reduces to

$$HD_2$$
 (low frequencies) =  $\frac{V_{in}}{4V_t}$  (5.31)

which is independent of the bias conditions. At high frequencies equation (5.29) reduces to

$$HD_2$$
 (low frequencies) =  $\frac{V_{in}}{8V_t}$  (5.32)

The third harmonic on node two is found by further elaborating equation (5.24). Using Table 4.1 together with the values of the second- and third-order nonlinearity coefficients from equations (3.14), (3.15), (3.22) and (3.24), the complex amplitude of the third harmonic is found to be

$$V_{2,3,0} = \frac{-V_{in}^3 g_m g_B^4 (g_B - 2g_\pi)}{24 (g_L + 3j\omega_1 C_L) V_t^2 (g_B + g_\pi)^5}$$
(5.33)

For a zero base resistance this reduces to

$$V_{2,3,0} = \frac{-V_{in}^3 g_m}{24 \left(g_L + 3j\omega_1 C_L\right) V_t^2}$$
 (5.34)

The third harmonic distortion is found by dividing  $V_{2,3,0}$  by the first-order response  $V_{2,1,0}$ . For a zero base resistance we find

$$HD_3 = \frac{V_{in}^2}{24V_t^2} \frac{g_L + j\omega_1 C_L}{g_L + 3j\omega_1 C_L}$$
 (5.35)

At low frequencies we have

$$HD_3$$
 (low frequencies) =  $\frac{V_{in}^2}{24V_t^2}$  (5.36)

and at high frequencies

$$HD_3 ext{ (high frequencies)} = \frac{V_{in}^2}{72V_t^2} ext{ (5.37)}$$

It is seen that the third harmonic distortion only depends on  $V_t$ , just as  $HD_2$ .

# 5.2.5 Simplifications

In Table 4.1 it is seen that many responses only require the knowledge of Volterra kernels for some of the arguments  $s_1$ ,  $s_2$  or  $s_3$  being equal. This means that a kernel does not need to be known completely. This can be exploited if one is only interested in the response of a given order at a given frequency, rather than the complete kernel. In that case, the expressions of the nonlinear current sources can be adapted to the response of interest.

Table 5.3 lists the nonlinear current sources for the computation of the second and third harmonics. They are derived from the values given in Table 5.1 and 5.2 by making the arguments of the lower-order Volterra kernels all equal to  $j\omega_1$ . Clearly, the expressions are much simpler than the expressions for the computation of the complete kernels. Table 5.4 gives the expressions for the calculation of third-order intermodulation products at  $2\omega_1 \pm \omega_2$ , which is found by putting  $s_1 = s_2 = j\omega_1$  and  $s_3 = \pm j\omega_2$  in Table 5.2.

type of basic nonlinearity	order	expression for nonlinear current source
one-dimensional	2	$K_{2g_1} \left( H_{1k}(j\omega_1) \right)^2$
(trans)conductance	3	$K_{3g_1} (H_{1k}(j\omega_1))^3 + 2K_{2g_1} H_{1k}(j\omega_1) H_{2k}(j\omega_1, j\omega_1)$
capacitance	2	$2j\omega_1 K_{2C_1} \left( H_{1k}(j\omega_1) \right)^2$
	3	$3j\omega_{1}K_{3C_{1}}\left(H_{1k}(j\omega_{1})\right)^{3} +6j\omega_{1}K_{2C_{1}}H_{1k}(j\omega_{1})H_{2k}(j\omega_{1},j\omega_{1})$
two-dimensional conductance (only cross-terms)	2	$K_{2_{g_1}\&g_2}H_{1k}(j\omega_1)H_{1l}(j\omega_1)$
	3	$K_{2g_{1}\&g_{2}}[H_{1k}(j\omega_{1})H_{2l}(j\omega_{1},j\omega_{1}) + H_{2k}(j\omega_{1},j\omega_{1})H_{1l}(j\omega_{1})]$ $+K_{32g_{1}\&g_{2}}(H_{1k}(j\omega_{1}))^{2}H_{1l}(j\omega_{1})$ $+K_{3g_{1}\&2g_{2}}H_{1k}(j\omega_{1})(H_{1l}(j\omega_{1}))^{2}$
three-dimensional conductance	2	0
(only cross-terms)	3	$K_{3_{g_1\&g_2\&g_3}}H_{1k}(j\omega_1)H_{1l}(j\omega_1)H_{1m}(j\omega_1)$

Table 5.3: Nonlinear second- and third-order current sources for the basic nonlinearities compute second and third harmonics of  $\omega_1$  with the Volterra series approach.

#### 5.2.6 Volterra kernels of currents

The nth-order Volterra kernel of a current through a nonlinear element consists of two components. The first component is the value of the nonlinear current source of order n that components to the given nonlinearity. However, a second component occurs which is caused by the nth-order kernel of the controlling voltage times the linear admittance. In other words, it is the current through the linearized element in the linearized circuit excited with the nonlinear current sources. For example, the second-order kernel of the collector current in Figure 5.2 can found by considering the linearized circuit of Figure 5.4: it is the sum of  $i_{NL2gm}$  with the current  $g_m H_{2_1}(s_1, s_2)$ .

	expression for	
type of basic	nonlinear current source	
nonlinearity	of order three	
(trans)canductance	$K_{3g_1}\left(H_{1k}(j\omega_1) ight)^2H_{1k}(\pm j\omega_2)$	
(trans)conductance	$+\frac{2}{3}K_{2g_1}\left[2H_{1k}(j\omega_1)H_{2k}(j\omega_1,\pm j\omega_2)+H_{1k}(\pm j\omega_2)H_{2k}(j\omega_1,j\omega_1)\right]$	
capacitance	$(2j\omega_1 \pm j\omega_2)K_{3_{C_1}} (H_{1k}(j\omega_1))^2 H_{1k}(\pm j\omega_2)$	
сараспансе	$+\frac{2}{3}(2j\omega_1 \pm j\omega_2)K_{2C_1}\left[2H_{1k}(j\omega_1)H_{2k}(j\omega_1, \pm j\omega_2) + H_{1k}(\pm j\omega_2)H_{2k}(j\omega_1, j\omega_1)\right]$	
	$rac{1}{3}K_{2_{g_1}\&g_2}igg[2H_{1k}(j\omega_1)H_{2l}(j\omega_1,\pm j\omega_2)+H_{1k}(\pm j\omega_2)H_{2l}(j\omega_1,j\omega_1)$	
two-dimensional	$+2H_{2k}(j\omega_{1},\pm j\omega_{2})H_{1l}(j\omega_{1})+H_{2k}(j\omega_{1},j\omega_{1})H_{1l}(\pm j\omega_{2})$	
conductance	$+\frac{1}{3}K_{3_{2g_1\&g_2}}\left[2H_{1k}(j\omega_1)H_{1k}(\pm j\omega_2)H_{1l}(j\omega_1) + (H_{1k}(j\omega_1))^2H_{1l}(\pm j\omega_2)\right]$	
(only cross-terms)	$+rac{1}{3}K_{3_{g_{1}\&2g_{2}}}igg[2H_{1k}(j\omega_{1})H_{1l}(j\omega_{1})H_{1l}(\pm j\omega_{2})+H_{1k}(\pm j\omega_{2})\left(H_{1l}(j\omega_{1}) ight)^{2}igg]$	
three-dimensional	$\frac{1}{3}K_{3_{g_1\&g_2\&g_3}} \left[ H_{1k}(j\omega_1)H_{1l}(\pm j\omega_2)H_{1m}(j\omega_1) \right]$	
conductance	٦ - ا	
(only cross-terms)	$+H_{1k}(j\omega_1)H_{1l}(j\omega_1)H_{1m}(\pm j\omega_2)+H_{1k}(\pm j\omega_2)H_{1l}(j\omega_1)H_{1m}(j\omega_1)$	

Table 5.4: Nonlinear third-order current sources for the basic nonlinearities to compute third-order intermodulation products at  $2\omega_1 \pm \omega_2$  with the Volterra series approach.

# 5.2.7 Interpretation of the results

Using the explained calculation method, the harmonics or intermodulation products of order p > 1 at the frequency  $|\pm m\omega_1 \pm n\omega_2|$  with m + n = p, are computed as a sum of contributions:

# nonlinearities
$$\sum_{k=1}^{m} i_{NL_{p_k}} \cdot TF_{i_{NL_{p_k}} \to output} \left( \pm m\omega_1 \pm n\omega_2 \right)$$
(5.38)

The sum is taken over all k nonlinearities in the network. In this equation,  $i_{NL_{p_k}}$  is the nonlinearity source of order p for nonlinearity k and  $TF_{i_{NL_{p_k}} \to output}(\pm m\omega_1 \pm n\omega_2)$  denotes the transfer function from the applied source to the output. The amplitudes of the external excitation have been taken equal to one here for simplicity.

The formulation of expression (5.38) makes reasoning about distortion possible. Indeed, transfer functions which determine the current sources on one hand, and the transfer function from the current sources to the circuit's output on the other hand, can be analyzed either nume cally or symbolically and interpreted. Moreover, since the current sources are applied in a line network, the effect of every current source, corresponding to one single nonlinearity, can be stu ied apart from the other ones, just as with a noise analysis. However, the different contribution of the nonlinear current sources are complex numbers, whereas in noise analysis the contril tions are positive real numbers. This means that distortion contributions can cancel, which is the case for contributions of noise sources. Moreover, the nonlinearities can interact, despite the one-to-one correspondence between a nonlinearity and a nonlinear current source. Consid for example, the third-order nonlinear current sources. Looking at Table 5.2, one can see t their value is determined by the second- and third-order nonlinearity coefficients  $K_2$  and  $K_3$ the nonlinearity under consideration, but also by the second-order response at the controlli voltage. This response is in turn determined by the effect of all second-order nonlinearities. interaction prevents the different contributions from being associated with only one nonlinear Hence, when the above method is used for symbolic analysis, such an interaction can complic the interpretation of the symbolic expressions. Fortunately, most contributions can be neglect in practical transistor circuits, so that after approximation very little contributions are retained the result and very often little interaction occurs. This will be illustrated with the examples Chapters 8.

#### 5.2.8 Factorization of the denominators

During the calculation of the second-order kernels in the circuit of Figure 5.2, the denomination of the second-order kernels could be neatly factorized, as seen for example in expression (5.1). This factorization can be generalized.

**Order 1** Consider first the solution of the linearized network. From equation (5.1) the vec  $\mathbf{H}_1(s)$  of linear transfer functions in the circuit is given by

$$\mathbf{H}_1(s) = \mathbf{Y}^{-1}.\mathbf{I}\mathbf{N}_1 = \frac{1}{\det(s)}\mathbf{Y}_N^{-1}.\mathbf{I}\mathbf{N}_1$$
 (5.2)

in which  $\mathbf{Y}_N^{-1}$  is the matrix formed by taking the numerator of every element of  $\mathbf{Y}^{-1}$ . It denominator of every element of  $\mathbf{Y}^{-1}$  is the determinant  $\det(s)$ . As a result, all first-ord transfer functions in a network have the same denominator. This means that this denomination only needs to be computed once.

Order 2 For order two, the linearized circuit is excited with the second-order nonlinear current sources. These are determined by the product of the first-order response of the controlling voltage at  $s_1$  with the same response at  $s_2$ . Hence, according to equation (5.39), the nonlinear current sources of order two have a common denominator  $(\det(s_1) \cdot \det(s_2))$ . The set of equations to be solved to find the second-order responses is then

$$\mathbf{Y}(s_1 + s_2).\mathbf{H}_2(s_1, s_2) = \frac{\mathbf{IN}_{2N}}{\det(s_1)\det(s_2)}$$
(5.40)

in which  $IN_{2N}$  is the numerator of the applied second-order nonlinear current sources and  $det(\cdot)$  is the determinant of the (C)MNA matrix. The solution of this matrix equation is

$$\mathbf{H}_{2}(s_{1}, s_{2}) = \frac{\mathbf{Y}_{N}^{-1}(s_{1} + s_{2}) \mathbf{I} \mathbf{N}_{2N}}{\det(s_{1} + s_{2}) \det(s_{1}) \det(s_{2})}$$
(5.41)

in which we used again the matrix  $\mathbf{Y}_N^{-1}$  that only contains elements without a denominator. It is seen that the second-order kernels have a common denominator that comprises products of the determinant of the (C)MNA matrix. This means that, once the determinant of the (C)MNA matrix is known, the denominator of the second-order Volterra kernels can be written down immediately. In other words, calculations of Volterra kernels can be limited to calculations of numerators.

If one is interested in second harmonics, then vector the second-order kernels in equation (5.41) has to be evaluated for  $s_1 = s_2$ . In this way, the denominator of all second harmonics in a circuit is given by

denominator of any 
$$H_{2k}(s_1, s_1) = (\det(s_1))^2 \det(2s_1)$$
 (5.42)

From equation (5.39) it is seen that the poles of the linearized circuit are the zeros of  $\det(s_1)$ . Then equation (5.42) reveals that a pole of the linearized circuit is a double pole for the second harmonic. In addition, there is an extra pole at half of the frequency of the original pole. Assume for example that a circuit has a dominant pole at a frequency  $f_1$ . This corresponds to a zero of the determinant of the (C)MNA matrix at  $f_1$ . On a Bode diagram the first-order response starts to decrease with 20 dB per decade from  $f_1$  on. The second-order response already decreases with 20 dB per decade from  $f_1/2$  and with 60 dB per decade from  $f_1$ .

Order 3 Factorization of denominators is also possible during calculations of Volterra kernels or responses of order three. An additional difficulty, however, is that the third-order nonlinear current sources consist of a component which contains the product of three first-order terms and a component that contains a second-order response. Hence the two components do not have the same numerator. If one wants to bring them to the same denominator, then the expressions for the nonlinear current sources of order three are quite complicated. However, considerable simplifications in the expressions of the nonlinear third-order current sources can be performed when only a specific third-order response is calculated.

Consider for example the calculation of the third-order intermodulation product at a frequency  $2\omega_1 + \omega_2$  at the output of a circuit that is excited by two sinusoidal input signals at frequencies  $\omega_1$  and  $\omega_2$ . The expression of the nonlinear current source of order three for a nonline (trans)conductance can be obtained from Table 5.4:

$$i_{NL3g_1} = K_{3g_1} (H_{1k}(j\omega_1))^2 H_{1k}(j\omega_2) + \frac{2}{3} K_{2g_1} \left[ 2H_{1k}(j\omega_1) H_{2k}(j\omega_1, j\omega_2) + H_{1k}(j\omega_2) H_{2k}(j\omega_1, j\omega_1) \right]$$
(5.4)

The Volterra kernels in this expression can be split in a numerator and a denominator. From equation (5.39) it is seen that a first-order kernel  $H_{1k}(j\omega_1)$  can be split as follows

$$H_{1k}(j\omega_1) = \frac{H_{1kN}(j\omega_1)}{\det(j\omega_1)}$$
(5.4)

Hereby  $H_{1kN}(j\omega_1)$  has no denominator. Using equation (5.41) a second-order kernel  $H_{2k}(j\omega_1, j\omega_1)$  can be written as

$$H_{2k}(j\omega_1, j\omega_2) = \frac{H_{2kN}(j\omega_1, j\omega_2)}{\det(j\omega_1 + j\omega_2)\det(j\omega_1)\det(j\omega_2)}$$
(5.4)

in which  $H_{2kN}(j\omega_1, j\omega_2)$  is a quantity without a denominator. Using equations (5.44) and (5.47) the nonlinear current source of order three for a nonlinear conductance, equation (5.43) become

$$i_{NL3g_{1}} = K_{3g_{1}} \left( \frac{H_{1kN}(j\omega_{1})}{\det(j\omega_{1})} \right)^{2} \cdot \frac{H_{1kN}(j\omega_{2})}{\det(j\omega_{2})} + \frac{2}{3} K_{2g_{1}} \left[ \frac{2H_{1kN}(j\omega_{1})}{\det(j\omega_{1})} \cdot \frac{H_{2kN}(j\omega_{1}, j\omega_{2})}{\det(j\omega_{1}) \det(j\omega_{2}) \det(j\omega_{1} + j\omega_{2})} + \frac{H_{1kN}(j\omega_{2})}{\det(j\omega_{2})} \cdot \frac{H_{2kN}(j\omega_{1}, j\omega_{1})}{(\det(j\omega_{1}))^{2} \det(2j\omega_{1})} \right]$$
(5.4)

The different terms can be brought to the same denominator. In this way, the common denominator of the nonlinear third-order current source becomes

denominator of 
$$i_{NL3g_1} = (\det(j\omega_1))^2 \det(j\omega_2) \det(j\omega_1 + j\omega_2) \det(2j\omega_1)$$
 (5.4)

and the numerator is given by

numerator of 
$$i_{NL3g_1} = K_{3g_1} (H_{1kN}(j\omega_1))^2 H_{1kN}(j\omega_2) \det(j\omega_1 + j\omega_2) \det(2j\omega_1)$$

$$+ \frac{2}{3} K_{2g_1} \left[ 2H_{1kN}(j\omega_1) H_{2kN}(j\omega_1, j\omega_2) \det(2j\omega_1) + H_{1kN}(j\omega_2) H_{2kN}(j\omega_1, j\omega_1) \det(j\omega_1 + j\omega_2) \right]$$
(5.48)

The nonlinear current sources of order three are applied to the linearized circuit at a frequency  $2j\omega_1+j\omega_2$ . As a result, the common denominator of the intermodulation products at  $2j\omega_1+j\omega_2$  everywhere in the circuit is given by the product of the denominator given in equation (5.47) and  $\det(2j\omega_1+j\omega_2)$ :

denominator of response at 
$$(2j\omega_1 + j\omega_2) = (\det(j\omega_1))^2 \det(j\omega_2) \det(j\omega_1 + j\omega_2) \det(2j\omega_1) \det(2j\omega_1 + j\omega_2)$$
 (5.49)

In a similar way, one finds that for the calculation of third harmonics at  $3\omega_1$  the common denominator of all nonlinear current sources of order three is given by

$$\det(2j\omega_1)\left(\det(j\omega_1)\right)^3\tag{5.50}$$

while the numerator of a third-order nonlinear current source corresponding to a one-dimensional conductance is given by

$$K_{3g_1} \left( H_{1kN}(j\omega_1) \right)^3 \det(2j\omega_1) + 2K_{2g_1} H_{1kN}(j\omega_1) H_{2kN}(j\omega_1, j\omega_1)$$
 (5.51)

## 5.3 Direct calculation of nonlinear responses

The Volterra kernels which have been used throughout this chapter have been defined for circuits with only one input port. For such circuits a Volterra kernel is an analytic function of several time or frequency variables. For multiple-input circuits, however, Volterra kernels become tensors [Chua 79a]. The second-order kernel of a voltage is a matrix of size ( $\#inputs \times \#inputs$ ), the third-order kernel is a ( $\#inputs \times \#inputs \times \#inputs$ ) tensor, and so on. Calculations with tensors are quite cumbersome. Although it is possible to get rid of tensor manipulations by the use of the Kronecker product [Sale 82], an extension of the above explained calculation method in terms of Volterra series for multiple-input systems is complex. However, the use of Volterra kernels can be circumvented to obtain the responses of interest in a multiple- input system. Indeed, in Appendix C a method is derived that immediately calculates the required responses (harmonics or intermodulation products). The method does not make use of tensors. Again the linearized network is solved repeatedly with different inputs. The difference with the Volterra series method for single-input circuits lies in the value of the nonlinear current sources. In this chapter the method is explained with an example circuit. Interested readers can go through the derivation of the method in Appendix C.

The method is explained with the example circuit of Figure 5.6. This circuit is a differential pair with bipolar transistors. By applying both a differential signal at the bases of the two transistors and a signal at the tail current source, a mixer operation is obtained. The two transistors  $Q_{1A}$  and  $Q_{1B}$  are each represented by a nonlinear transconductance that corresponds to the nonlinear dependence of the collector current on the base-emitter voltage. The two transistors match. Therefore, their transconductance and the corresponding higher-order coefficients will be represented by the same symbol in the subsequent computations.

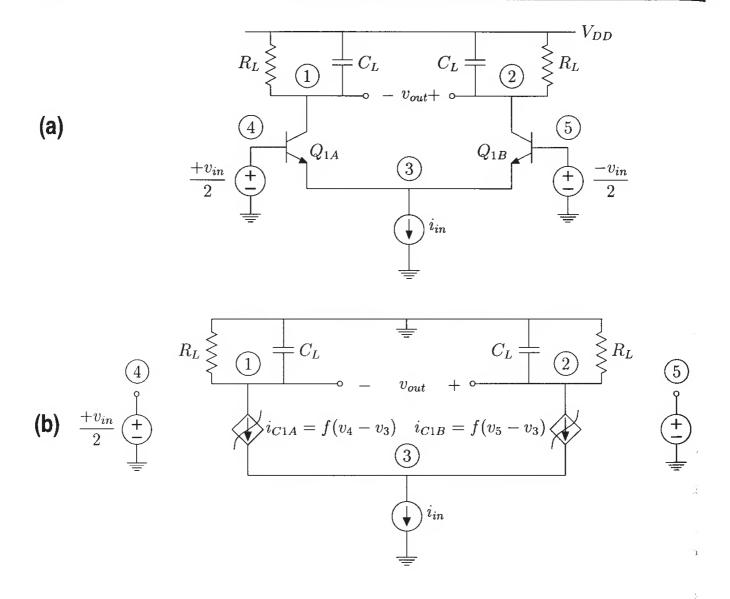


Figure 5.6: A differential pair used as a simple mixer (a) and its AC-equivalent circuit (b).

The circuit is excited at two different input ports by two sinusoidal signals  $i_{in}(t)$  and  $v_{in}(t)$  at frequencies  $\omega_1$  and  $\omega_2$ , respectively:

$$v_{in}(t) = \operatorname{Re}\left(V_{in}e^{j\omega_1t}\right)$$
 (5.5)  
 $i_{in}(t) = \operatorname{Re}\left(I_{in}e^{j\omega_2t}\right)$  (5.5)

Under steady-state conditions, every node voltage  $v_x(t)$  consists of a sum of harmonic fund

tions:

$$\begin{aligned} v_{x}(t) &= \\ & \operatorname{Re}(V_{x,1,0} e^{j\omega_{1}t}) + \operatorname{Re}(V_{x,0,1} e^{j\omega_{2}t}) & \operatorname{linear \ response} \\ & + \operatorname{Re}(V_{x,2,0} e^{j2\omega_{1}t}) + \operatorname{Re}(V_{x,0,2} e^{j2\omega_{2}t}) \\ & + \operatorname{Re}(V_{x,1,1} e^{j(\omega_{1}+\omega_{2})t}) + \operatorname{Re}(V_{x,1,-1} e^{j(\omega_{1}-\omega_{2})t}) \end{aligned} \right\} \quad 2\operatorname{nd-order \ response}$$
 
$$+ \operatorname{Re}(V_{x,3,0} e^{j3\omega_{1}t}) + \operatorname{Re}(V_{x,0,3} e^{j3\omega_{2}t}) + \dots$$
 
$$3\operatorname{rd-order \ response}$$
 
$$+ \dots$$
 
$$\vdots$$

In this equation,  $V_{x,m,n}$  is the phasor of the component of the voltage at node x at the frequency  $m\omega_1 + n\omega_2$ . Similarly, the current through a branch k can be written using phasors of the form  $I_{k,m,n}$ .

Equation (5.54) can be rewritten as

$$v_x(t) = \sum_{m=0}^{+\infty} \sum_{n=0}^{+\infty} \operatorname{Re}\left(V_{x,m,n} e^{j(m\omega_1 + n\omega_2)t}\right)$$
 (5.55)

The goal of this section is to explain a method to compute the complex phasors  $V_{x,m,n}$ .

Since the real part of a complex number is half of the sum of this number and its complex conjugate, equation (5.55) can also be written as

$$v_x(t) = \frac{1}{2} \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{x,m,n} e^{j(m\omega_1 + n\omega_2)t}$$
 (5.56)

in which

$$V_{x,m,n} = V_{x,-m,-n}^* (5.57)$$

In other words, the response at  $m\omega_1 + n\omega_2$  is the complex conjugate of the response at  $-m\omega_1 - n\omega_2$ .

# 5.3.1 First-order responses

First, the responses to  $v_{in}$  only are computed in the linearized circuit. This is performed with the matrix equation

$$\mathbf{Y}(j\omega_1).\mathbf{U}_{1,0} = \mathbf{IN}_{1,1,0} \tag{5.58}$$

In this equation Y is the (C)MNA matrix,  $U_{1,0}$  is the vector of the phasors  $V_{i,1,0}$  (and possibly  $I_{k,1,0}$ ) and  $IN_{1,1,0}$  is a vector whose only nonzero components are the terms in the network equations that contain  $V_{in}$ . The circuit output and the voltages that control a nonlinearity can be written as a linear combination of the components of vector  $U_{1,0}$ .

Next the responses in the linearized network to  $I_{in}$  only are computed with the equation

$$\mathbf{Y}(j\omega_2).\mathbf{U}_{0,1} = \mathbf{IN}_{1,0,1}$$
 (5.59)

in which  $U_{0,1}$  is the vector of the phasors  $V_{i,0,1}$  (and possibly  $I_{k,0,1}$ ) and  $IN_{1,0,1}$  is a vector whos only nonzero components are the terms in the network equations that contain  $I_{in}$ . The circuit output and the voltages that control a nonlinearity can be written as a linear combination of the components of vector  $U_{0,1}$ .

For the example circuit of Figure 5.6, the computation of the first-order responses is illustrated in Figure 5.7a and 5.7b.

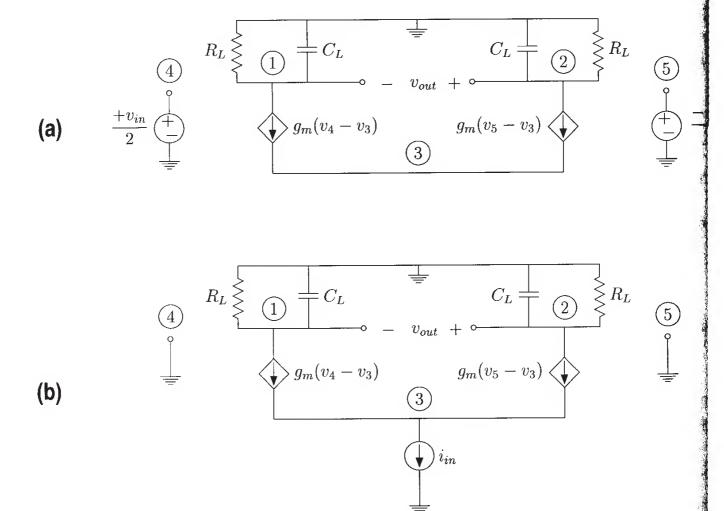


Figure 5.7: Calculation of the first-order responses to  $v_{in}$  only (a) and  $i_{in}$  only (b), respectively in the circuit of Figure 5.6.

We begin by computing the first-order responses to  $v_{in}$  only. This is performed with the cuit in Figure 5.7a. By applying Kirchoff's current law at the nodes 1, 2 and 3, the matrix

equation (5.58) can be set up. This yields

$$\begin{bmatrix} g_L + j\omega_1 C_L & 0 & -g_m \\ 0 & g_L + j\omega_1 C_L & -g_m \\ 0 & 0 & 2g_m \end{bmatrix} \cdot \begin{bmatrix} V_{1,1,0} \\ V_{2,1,0} \\ V_{3,1,0} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}g_m V_{in} \\ \frac{1}{2}g_m V_{in} \\ 0 \end{bmatrix}$$
(5.60)

The first row in this matrix equation corresponds to the current law formulation at node 1, the second row to node 2 and the third one to node 3. Note that again the CMNA formulation has been used to write down the network equations. The determinant of the CMNA matrix is found from equation (5.60):

$$\det(j\omega_1) = \begin{vmatrix} g_L + j\omega_1 C_L & 0 & -g_m \\ 0 & g_L + j\omega_1 C_L & -g_m \\ 0 & 0 & 2g_m \end{vmatrix} = 2g_m (g_L + j\omega_1 C_L)^2$$
 (5.61)

The responses to  $V_{in}$  only can be found by using the rule of Cramer. For the response at node 1 we find in this way

$$V_{1,1,0} = rac{1}{\det(j\omega_1)} egin{array}{cccc} -rac{1}{2}g_mV_{in} & 0 & -g_m \ rac{1}{2}g_mV_{in} & g_L + j\omega_1C_L & -g_m \ 0 & 0 & 2g_m \end{array}$$

$$= \frac{-g_m}{2(g_L + j\omega_1 C_L)} V_{in} \tag{5.62}$$

Similarly, the response at node 2 is found to be

$$V_{2,1,0} = \frac{g_m}{2(q_L + i\omega_1 C_L)} V_{in}$$
(5.63)

which is the opposite of  $V_{1,1,0}$ . Further we find that

$$V_{3,1,0} = 0 (5.64)$$

 $V_{3,1,0}$  is zero since for differential signals applied to the base of the two transistors, the common emitter point is an AC ground. The output signal is the difference of  $V_{1,1,0}$  and  $V_{2,1,0}$ :

$$V_{out,1,0} = V_{2,1,0} - V_{1,1,0} (5.65)$$

Using equations (5.62) and (5.63) we find

$$V_{out,1,0} = \frac{g_m}{g_L + j\omega_1 C_L} V_{in}$$
 (5.66)

In terms of the determinant of the CMNA matrix this can be rewritten as

$$V_{out,1,0} = \frac{2g_m^2 (g_L + j\omega_1 C_L)}{\det(j\omega_1)} V_{in}$$
(5.67)

In the Volterra series approach that has been explained in Section 5.2 nonlinear current sources were applied in order to compute the higher-order Volterra kernels. These current source were determined by the lower-order Volterra kernels of the voltages that control the different nonlinearities in the circuit. With the direct calculation of nonlinear responses we will follow a similar approach. Again, the value of the nonlinear current sources for the computation of higher-order responses will depend on first-order responses of the voltages that control the nonlinearities. Hence we have compute these first-order responses.

The nonlinearities in this circuit are the transconductances of the two transistors. The transconductance that corresponds to transistor  $Q_{1A}$  is controlled by the voltage difference between nodes 4 and 3 and the controlling voltage for the transconductance corresponding to  $Q_{1B}$  the voltage difference between nodes 5 and 3:

$$V_{43,1,0} = V_{4,1,0} - V_{3,1,0} (5.6)$$

$$V_{53,1,0} = V_{5,1,0} - V_{3,1,0} (5.6)$$

From Figure 5.7a it is seen that  $V_{4,1,0}$  and  $V_{5,1,0}$  are  $+V_{in}/2$  and  $-V_{in}/2$ , respectively. Since  $V_{3,1}$  is zero we find

$$V_{43,1,0} = V_{in}/2 (5.76$$

$$V_{53,1,0} = -V_{in}/2 (5.7$$

Next, the responses to  $I_{in}$  are computed. Applying Kirchoff's current law at the nodes 1, and 3 in Figure 5.7b, yields the following matrix equation

$$\begin{bmatrix} g_L + j\omega_2 C_L & 0 & -g_m \\ 0 & g_L + j\omega_2 C_L & -g_m \\ 0 & 0 & 2g_m \end{bmatrix} \cdot \begin{bmatrix} V_{1,0,1} \\ V_{2,0,1} \\ V_{3,0,1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -I_{in} \end{bmatrix}$$
 (5.7)

Applying the rule of Cramer yields the response to  $i_{in}$  at node 1:

$$V_{1,0,1} = \frac{1}{\det(j\omega_2)} \begin{vmatrix} 0 & 0 & -g_m \\ 0 & g_L + j\omega_2 C_L & -g_m \\ -I_{in} & 0 & 2g_m \end{vmatrix}$$

$$= -\frac{1}{2(g_L + j\omega_2 C_L)} I_{in}$$
(5.7)

where  $\det(j\omega_2)$  is found from equation (5.61) by replacing  $\omega_1$  with  $\omega_2$ . The response at node is identical to  $V_{1,0,1}$ :

$$V_{2,0,1} = -\frac{1}{2(q_L + j\omega_2 C_L)} I_{in}$$
(5.7)

The response at node 3 is found as

$$V_{3,0,1} = \frac{1}{\det(j\omega_2)} \begin{vmatrix} g_L + j\omega_2 C_L & 0 & 0\\ 0 & g_L + j\omega_2 C_L & 0\\ 0 & 0 & -I_{in} \end{vmatrix}$$
$$= -\frac{1}{2g_m} I_{in}$$
 (5.75)

The output voltage is found from equations (5.73) and (5.74):

$$V_{out,0,1} = V_{2,0,1} - V_{1,0,1} = \frac{-2g_m (g_L + j\omega_2 C_L)}{\det(j\omega_2)} I_{in} = -\frac{1}{g_L + j\omega_2 C_L} I_{in}$$
 (5.76)

Finally, we determine the voltages that control the two nonlinear transconductances:

$$V_{43,0,1} = V_{4,0,1} - V_{3,0,1} (5.77)$$

$$V_{53,0,1} = V_{5,0,1} - V_{3,0,1} (5.78)$$

From Figure 5.7b it is seen that both  $V_{4,0,1}$  and  $V_{5,0,1}$  are zero. Using equation (5.75) we find

$$V_{43,0,1} = V_{53,0,1} = \frac{1}{2q_m} I_{in} \tag{5.79}$$

### 5.3.2 Second-order responses

With the information from the previous step the second-order responses can be computed. From Section 2.4 and 4.4 we know that second-order behavior gives rise to responses at 0Hz,  $2\omega_1$ ,  $2\omega_2$  and  $|\omega_1 \pm \omega_2|$ . For the calculation of each of these responses, again the same linearized network must be solved as the one that has been used to compute the first-order responses. However, the inputs are different now: instead of the external excitations, so-called *nonlinear current sources* of order two must be applied. There is one current source for each basic nonlinearity in the circuit and the source is placed in parallel with the linearized equivalent of the nonlinearity. Clearly, this approach is similar to the Volterra series approach described in Section 5.2. However, the value of the nonlinear current source differs for the two approaches. For the different basic nonlinearities the expressions of the nonlinear current sources of order two that have to be applied to directly compute the responses at  $2\omega_1$  and  $|\omega_1 \pm \omega_2|$ , are listed in Table 5.5. The expressions for the nonlinear current sources of order two for the computation of responses at  $2\omega_2$  are similar to the expressions for the computation of responses at  $2\omega_1$ 

Since the responses at  $2\omega_1$  and  $2\omega_2$  are determined by one single excitation, it is clear that the computation of these responses can also be performed with the Volterra series method for single-input systems that has been explained in Section 5.2. However, the second-order intermodulation product cannot be computed with that method since it is determined by the mixing of two signals from a different input port. It is however possible to use Volterra series as well, with the extra burden that we need to introduce tensors, as already mentioned above.

type of nonlinearity	nonlinear current source	nonlinear current source
урс	for response at $ \omega_1 \pm \omega_2 $	for response at $2\omega_1$
(trans)conductance	$K_{2g_1} V_{i,1,0} V_{i,0,\pm 1}$	$\frac{K_{2g_1}}{2} (V_{i,1,0})^2$
capacitor	$\int j(\omega_1 \pm \omega_2) K_{2C_1} V_{i,1,0} V_{i,0,\pm 1}$	$j\omega_1  K_{2C_1}  (V_{i,1,0})^2$
two-dimensional	$\frac{K_{2_{g_1} \& g_2}}{2}  V_{i,1,0} V_{j,0,\pm 1}$	7.5
conductance		$\frac{K_{2g_1\&g_2}}{2} V_{i,1,0} V_{j,1,0}$
(only cross-terms)	$+\frac{K_{2g_1\&g_2}}{2}V_{i,0,\pm 1}V_{j,1,0}$	

Table 5.5: Nonlinear second-order current sources for the basic nonlinearities to directly compute second-order intermodulation products at  $|\omega_1 \pm \omega_2|$  and second harmonics at  $2\omega_1$ . The controlling voltages are  $v_i$  for the nonlinear (trans)conductance and the nonlinear capacitor and  $v_i$  for the two-dimensional conductance.

The interpretation of the values of the nonlinear current sources from Table 5.5 is as follows the second-order nonlinearity of every one-dimensional conductance or capacitance combine the first-order response of its controlling voltage due to  $V_{in}$  only, with the first-order response of its controlling voltage due to  $I_{in}$  only, to a second-order signal. This signal, at frequence  $|\omega_1 \pm \omega_2|$ , then propagates through the rest of the circuit. When considering this propagation only the linearized elements need to be taken into account. Indeed, any interaction of the second order signal with another one yields a response of order higher than two.

For the computation of the responses at the frequency  $|\omega_1 \pm \omega_2|$ , the following set of equation has to be solved:

$$\mathbf{Y}(j\omega_1 \pm j\omega_2).\mathbf{U}_{1,1} = \mathbf{IN}_{2,1,1}$$
 (5.80)

Here  $U_{1,1}$  is the vector of phasors of the components at  $|\omega_1 \pm \omega_2|$  of the node voltages and possible of some branch currents as well. The matrix  $\mathbf{Y}(j\omega_1 \pm j\omega_2)$  is again the (C)MNA matrix of the circuit, now evaluated at the frequency  $|\omega_1 \pm \omega_2|$ . The right-hand side  $\mathbf{IN}_{2,1,1}$  is a vector the contains the nonlinear current sources of order two, in particular the sources for the computation of the responses at  $|\omega_1 \pm \omega_2|$ , which can be found in Table 5.5.

For the computation of the responses at  $2\omega_1$  the following set of equations needs to be solved

$$Y(2j\omega_1).U_{2,0} = IN_{2,2,0}$$
 (5.81)

where  $U_{2,0}$  now contains phasors of responses at  $2\omega_1$  and  $IN_{2,2,0}$  is the vector of nonlinear current sources of order two, in particular the sources for the computation of the responses

 $2\omega_1$ , which can be found in the rightmost column of Table 5.5. Similarly, the computation of responses at  $2\omega_2$  is performed with the following set of equations:

$$\mathbf{Y}(2j\omega_2).\mathbf{U}_{0,2} = \mathbf{IN}_{2,0,2}$$
 (5.82)

It is seen that with the direct computation of the wanted responses each time a new set of equations needs to be solved whenever an additional response is required. This was not the case with the Volterra series approach.

In order to illustrate the method, we will now compute the second-order intermodulation product at  $\omega_1 + \omega_2$  at the output of the example circuit of Figure 5.2. To this purpose, appropriate nonlinear current sources of order two are applied to the linear circuit of Figure 5.3. The resulting network that has to be solved is shown in Figure 5.8.

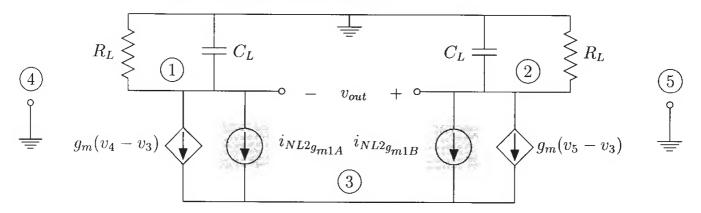


Figure 5.8: Equivalent circuit that has to be solved for the computation of the second-order intermodulation products in the circuit of Figure 5.6.

First, we set up the matrix equation (5.80). To this purpose, the current law of Kirchoff is applied at the nodes 1, 2 and 3. This yields

$$\begin{bmatrix} g_{L} + j(\omega_{1} + \omega_{2})C_{L} & 0 & -g_{m} \\ 0 & g_{L} + j(\omega_{1} + \omega_{2})C_{L} & -g_{m} \\ 0 & 0 & 2g_{m} \end{bmatrix} \cdot \begin{bmatrix} V_{1,1,1} \\ V_{2,1,1} \\ V_{3,1,1} \end{bmatrix} = \begin{bmatrix} -i_{NL2}g_{m1A} \\ -i_{NL2}g_{m1B} \\ i_{NL2}g_{m1A} + i_{NL2}g_{m1B} \end{bmatrix}$$
(5.83)

The right-hand side contains the nonlinear current sources  $i_{NL2g_{m1A}}$  and  $i_{NL2g_{m1B}}$  that correspond to the transconductance of transistors  $Q_{1A}$  and  $Q_{1B}$ , respectively. Their value can be obtained using Table 5.5. For  $i_{NL2g_{m1A}}$  we find

$$i_{NL2g_{m1A}} = K_{2g_m} V_{43,1,0} V_{43,0,1} (5.84)$$

Using equations (5.70) and (5.79) this becomes

$$i_{NL2g_{m1A}} = \frac{K_{2g_m}}{4g_m} V_{in} I_{in} \tag{5.85}$$

The nonlinear current source  $i_{NL2g_{m1B}}$  is given by

$$i_{NL2g_{m1B}} = K_{2g_m} V_{53,1,0} V_{53,0,1} (5.86)$$

From equations (5.71) and (5.79) we find

$$i_{NL2g_{m1B}} = -\frac{K_{2g_m}}{4g_m} V_{in} I_{in} (5.87)$$

which is the opposite of  $i_{NL2g_{m1A}}$ .

Now the phasors of the intermodulation products at the three nodes can be computed. Using Cramer's rule with the matrix equation (5.83) we find

$$V_{1,1,1} = \frac{g_m (g_L + j (\omega_1 + \omega_2) C_L) \left( -i_{NL2} g_{m1A} + i_{NL2} g_{m1B} \right)}{\det(j (\omega_1 + \omega_2))}$$
(5.88)

$$V_{2,1,1} = \frac{g_m \left(g_L + j \left(\omega_1 + \omega_2\right) C_L\right) \left(i_{NL2g_{m1A}} - i_{NL2g_{m1B}}\right)}{\det(j \left(\omega_1 + \omega_2\right))}$$
(5.89)

$$V_{3,1,1} = \frac{(g_L + j(\omega_1 + \omega_2)C_L)^2 (i_{NL2g_{m1A}} + i_{NL2g_{m1B}})}{\det(j(\omega_1 + \omega_2))}$$
(5.96)

Since  $i_{NL2g_{m1B}} = -i_{NL2g_{m1A}}$  we find

$$V_{3+1} = 0 ag{5.91}$$

This means that the second-order response at  $\omega_1 + \omega_2$  at the common-emitter point is zero.

The second-order intermodulation product at the output is the difference of the intermodulation products at nodes 1 and 2:

$$V_{out,1,1} = V_{2,1,1} - V_{1,1,1} (5.9)$$

Using equations (5.85), (5.87), (5.88) and (5.89) this becomes

$$V_{out,1,1} = \frac{K_{2g_m} V_{in} I_{in}}{2g_m \left(g_L + j \left(\omega_1 + \omega_2\right) C_L\right)}$$
(5.92)

The expression for the response at  $\omega_1 + \omega_2$  can be interpreted as follows. Assume that the mixing circuit of Figure 5.6 is used as an upconverter. In this case,  $V_{out,1,1}$  is the wanted mixing product it is seen to depend simultaneously on  $V_{in}$  and  $I_{in}$ . Suppose that the baseband signal is applicant the bases of transistors  $Q_{1A}$  and  $Q_{1B}$ . The AC current delivered by the current source at the common emitter is proportional to the local oscillator signal. Then the conversion gain of the simple mixer can be found by dividing  $V_{out,1,1}$  in equation (5.93) by the amplitude  $V_{in}$  of the baseband signal. This yields

Conversion gain = 
$$\frac{K_{2g_m}I_{in}}{2g_m(g_L + j(\omega_1 + \omega_2)C_L)} = \frac{K'_{2g_m}I_{in}}{2(g_L + j(\omega_1 + \omega_2)C_L)}$$
 (5.94)

It is seen that the conversion gain is proportional to the local oscillator signal. This is only true of course if the circuit operates in a weakly nonlinear way. This is not the case with all mixer configurations, as already mentioned in Section 2.5. Further, it is seen that the conversion gain is proportional to the normalized second-order nonlinearity coefficient of the collector current nonlinearity of the two transistors. Finally, it is seen that the conversion gain is higher when the impedance at the output is higher.

Assuming that the collector current satisfies the simple exponential relationship of equation (3.12), then the second-order nonlinearity coefficient  $K'_{2g_m}$  is equal to  $1/(2V_t)$  (see equation (3.14)), and the conversion gain reduces to

Conversion gain = 
$$\frac{I_{in}}{4V_t (g_L + j (\omega_1 + \omega_2) C_L)}$$
 (5.95)

If the mixer is used as a downconverter, then the response of interest is at the difference frequency  $|\omega_1 - \omega_2|$ . The conversion gain in this case is found by replacing in equation (5.95)  $\omega_2$  with  $-\omega_2$ . In this way we obtain

Conversion gain (difference frequency) = 
$$\frac{I_{in}}{4V_t \left(g_L + j \left(\omega_1 - \omega_2\right) C_L\right)}$$
 (5.96)

Compared to the upconversion situation, it is seen that the load impedance is evaluated now at the difference frequency. In the case of a capacitive load, this results in a higher conversion gain.

## 5.3.3 Third-order and higher-order responses

We know from previous chapters that third-order nonlinear behavior in a circuit that is excited with two sinusoidal signals at  $\omega_1$  and  $\omega_2$ , gives rise to responses at  $\omega_1$ ,  $\omega_2$ ,  $|2\omega_1 \pm \omega_2|$ ,  $|2\omega_2 \pm \omega_1|$ ,  $3\omega_1$  and  $3\omega_2$ . In Appendix C it is shown that these responses can be computed in a similar way to the computation of the second-order responses: the responses are found by solving the linearized network that is excited by nonlinear current sources which are now of order three. The value depends on the type of nonlinearity and on lower-order responses. As an example, the nonlinear current sources for the computation of the intermodulation products at  $|2\omega_1 \pm \omega_2|$  are shown in Table 5.6. The nonlinear current sources to compute intermodulation products at  $|2\omega_2 \pm \omega_1|$  can be derived from this table by changing the role of  $\omega_1$  and  $\omega_2$ .

The values of the sources from Table 5.6 can be reconstructed by considering all possibilities to produce a third-order signal out of the lower-order responses at the controlling voltages. Consider for example a one-dimensional nonlinear conductance. Its second-order nonlinearity combines the response at its controlling voltage at  $\omega_1$  with the second-order response at  $\omega_1 + \omega_2$ , giving rise to the first term of the expression of the third-order nonlinear current source. The second term is due to the second-order nonlinearity that acts upon the first-order response at  $\omega_1$  and the second-order response at  $2\omega_1$ . Finally, the third term is caused by the third-order nonlinearity that takes the first-order response at  $\omega_2$ .

Other third-order responses like harmonics, desensitizations, third-order compressions or expansions can be computed similarly. The only difference is in the expressions of the nonlinear

type of nonlinearity	nonlinear current source for response at $2\omega_1\pm\omega_2$	
(trans)conductance	$K_{2g_1}V_{i,1,0}V_{i,1,\pm 1} + K_{2g_1}V_{i,0,\pm 1}V_{i,2,0} + \frac{3}{4}K_{3g_1}V_{i,1,0}^2V_{i,0,\pm 1}$	
capacitor	$\left[ (2j\omega_1 \pm j\omega_2) \left[ K_{2C_1} V_{i,1,0} V_{i,1,\pm 1} + K_{2C_1} V_{i,0,\pm 1} V_{i,2,0} + \frac{3}{4} K_{3C_1} V_{i,1,0}^2 V_{i,0,\pm 1} \right] \right]$	
two-dimensional	$\frac{1}{2}K_{2g_1\&g_2}\left[V_{i,0,\pm 1}V_{j,2,0} + V_{i,1,0}V_{j,1,\pm 1} + V_{i,1,\pm 1}V_{j,1,0} + V_{i,2,0}V_{j,0,\pm 1}\right]$	
conductance	$+\frac{1}{4}K_{3_{2g_1\&g_2}}\left[2V_{i,0,\pm 1}V_{i,1,0}V_{j,1,0}+V_{i,1,0}^2V_{j,0,\pm 1}\right]$	
(only cross-terms)	$+\frac{1}{4}K_{3_{g_1\&2g_2}}\left[2V_{i,1,0}V_{j,0,\pm 1}V_{j,1,0}+V_{i,0,\pm 1}V_{j,1,0}^2\right]$	
three-dimensional		
conductance	$\left  \frac{1}{4} K_{3g_1 \& g_2 \& g_3} \left[ V_{i,0,\pm 1} V_{j,1,0} V_{k,1,0} + V_{i,1,0} V_{j,0,\pm 1} V_{k,1,0} + V_{i,1,0} V_{j,1,0} V_{k,0,\pm 1} \right] \right  $	
(only cross-terms)		

Table 5.6: Nonlinear third-order current sources for the basic nonlinearities to compute third-order intermodulation products at  $|2\omega_1 \pm \omega_2|$ . The controlling voltages are  $v_i$  for the nonlinear (trans)conductance and the nonlinear capacitor,  $v_i$  and  $v_j$  for the two-dimensional conductance and  $v_i$ ,  $v_j$  and  $v_k$  for the three-dimensional conductance.

current sources. Table 5.7 lists the nonlinear current sources for the computation of the third harmonic of  $\omega_2$ .

The computation of responses of order higher than three is seldom of interest in weakly nonlinear circuits. If required, these responses can be computed in a similar way. The higher-order nonlinear current sources will again be dependent on lower-order responses.

Let us return now to the example circuit of Figure 5.6 which is used as a simple mixer. It is assumed that the mixer is used as an upconverter. The baseband signal is a sinusoidal voltage with frequency  $\omega_1$  and the local oscillator signal has a frequency  $\omega_2$ . We first consider the situation where the baseband signal is applied at the base of the two transistors  $Q_{1A}$  and  $Q_{1B}$  and the local oscillator signal is proportional to the current of the source at the common emitter. The intermodulation product at the output at the frequency  $2\omega_1 + \omega_2$  is a signal that is very close to the wanted output signal at  $\omega_1 + \omega_2$ . Hence it is important to know this unwanted intermodulation product, which can be considered as in-band distortion.

For the computation of third-order intermodulation products at  $2\omega_1 + \omega_2$  the linearized circuits excited with nonlinear current sources of order three, as shown in Figure 5.9.

type of nonlinearity	nonlinear current source for response at $3\omega_2$	
(trans)conductance	$K_{2g_1}V_{i,0,1}V_{i,0,2} + \frac{1}{4}K_{3g_1}V_{i,0,1}^3$	
capacitor	$3j\omega_{2}\bigg[K_{2_{C_{1}}}V_{i,0,1}V_{i,0,2} + \frac{1}{4}K_{3_{C_{1}}}V_{i,0,1}^{3}\bigg]$	
two-dimensional	$ \frac{1}{2}K_{2g_1\&g_2} \left[ V_{i,0,1}V_{j,0,2} + V_{i,0,2}V_{j,0,1} \right]  + \frac{1}{4}K_{32g_1\&g_2}V_{i,0,1}^2V_{j,0,1} + \frac{1}{4}K_{3g_1\&2g_2}V_{i,0,1}V_{j,0,1}^2 $	
conductance		
(only cross-terms)		
three-dimensional		
conductance	$\frac{1}{4}K_{3_{g_1}\&g_2\&g_3}V_{i,0,1}V_{j,0,1}V_{k,0,1}$	
(only cross-terms)		

Table 5.7: Nonlinear third-order current sources for the basic nonlinearities to compute the third harmonic at  $3\omega_2$ . The controlling voltages are  $v_i$  for the nonlinear (trans)conductance and the nonlinear capacitor,  $v_i$  and  $v_j$  for the two-dimensional conductance and  $v_i$ ,  $v_j$  and  $v_k$  for the three-dimensional conductance.

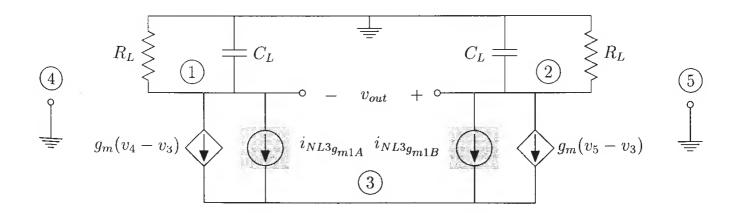


Figure 5.9: Equivalent circuit that has to be solved for the computation of third-order intermodulation products in the circuit of Figure 5.6.

The value of the nonlinear current sources  $i_{NL3g_{m1A}}$  and  $i_{NL3g_{m1B}}$  is obtained from Table 5.6.

For  $i_{NL3g_{m1A}}$  we find

$$i_{NL3g_{m1A}} = K_{2g_m} V_{43,1,0} V_{43,1,1} + K_{2g_m} V_{43,0,1} V_{43,2,0} + \frac{3}{4} K_{3g_m} V_{43,1,0}^2 V_{43,0,1}$$
 (5.97)

The phasors  $V_{43,1,0}$  and  $V_{43,0,1}$  in this equation have already been obtained in equations (5.70) and (5.79), respectively. The phasor of the second-order response of the controlling voltage,  $V_{43,1,1}$  is found as

$$V_{43,1,1} = V_{4,1,1} - V_{3,1,1} \tag{5.98}$$

The voltage at node 4 is fixed to  $v_{in}/2$ . Since the input voltage is assumed to be a pure sine wave, no higher-order harmonic or intermodulation products are present at node 4. Further, we know from equation (5.91) that  $V_{3,1,1}$  is zero. Hence

$$V_{43,1,1} = 0 (5.99)$$

The second term in the expression of  $i_{NL3g_{m1A}}$  contains the phasor  $V_{43,2,0}$ , which is the second harmonic of the controlling voltage at  $2\omega_1$ :

$$V_{43,2,0} = V_{4,2,0} - V_{3,2,0} (5.100)$$

Here  $V_{4,2,0}$  is zero since node 4 is fixed to  $v_{in}/2$ . The phasor  $V_{3,2,0}$  can be computed in the same way as  $V_{3,1,1}$  has been computed in equation (5.90):

$$V_{3,2,0} = \frac{(g_L + 2j\omega_1 C_L)^2 \left(i_{NL2g_{m1A}} + i_{NL2g_{m1B}}\right)}{\det(2j\omega_1)}$$
(5.101)

in which the nonlinear current sources now correspond to the calculation of second harmonics  $\omega_1$  (see rightmost column of Table 5.5). In this way, we find

$$V_{3,2,0} = \frac{K_{2g_m}}{8g_m} V_{in}^2 \tag{5.102}$$

and hence

$$V_{43,2,0} = -\frac{K_{2g_m}}{8g_m}V_{in}^2 \tag{5.103}$$

Using equations (5.70), (5.79), (5.99) and (5.103) the nonlinear current source of order thres  $i_{NL3g_{m1A}}$  given in equation (5.97) now becomes

$$i_{NL3g_{m1A}} = -\frac{1}{16} \frac{K_{2g_m}^2}{g_m^2} V_{in}^2 I_{in} + \frac{3}{32} \frac{K_{3g_m}}{g_m} V_{in}^2 I_{in}$$
(5.104)

The third-order nonlinear current source  $i_{NL3g_{m1B}}$  that corresponds to transistor  $Q_{1B}$  is given by

$$i_{NL3g_{m1B}} = K_{2g_m} V_{53,1,0} V_{53,1,1} + K_{2g_m} V_{53,0,1} V_{53,2,0} + \frac{3}{4} K_{3g_m} V_{53,1,0}^2 V_{53,0,1}$$
 (5.105)

Similarly to equation (5.99) it is found that

$$V_{53,1,1} = 0 (5.106)$$

Further, equations (5.70) and (5.71) reveal that  $V_{53,1,0} = -V_{43,1,0}$ . Next, it is found that

$$V_{53,2,0} = V_{43,2,0} (5.107)$$

since both  $V_{5,2,0}$  and  $V_{4,2,0}$  are zero. Further we know that  $V_{53,0,1}=V_{43,0,1}$ . As a result we find that

$$i_{NL3g_{m1B}} = i_{NL3g_{m1A}} (5.108)$$

We now have the necessary data to compute the third-order intermodulation product at  $2\omega_1$  +  $\omega_2$ . Applying Kirchoff's law in the circuit of Figure 5.9 yields the following set of equations

$$\begin{bmatrix} g_{L} + j(2\omega_{1} + \omega_{2})C_{L} & 0 & -g_{m} \\ 0 & g_{L} + j(2\omega_{1} + \omega_{2})C_{L} & -g_{m} \\ 0 & 0 & 2g_{m} \end{bmatrix} \cdot \begin{bmatrix} V_{1,2,1} \\ V_{2,2,1} \\ V_{3,2,1} \end{bmatrix}$$

$$= \begin{bmatrix} -i_{NL3}g_{m1A} \\ -i_{NL3}g_{m1B} \\ i_{NL3}g_{m1A} + i_{NL3}g_{m1B} \end{bmatrix} (5.109)$$

From this equation we find with the rule of Cramer

$$V_{1,2,1} = \frac{g_m \left(g_L + j \left(2\omega_1 + \omega_2\right) C_L\right) \left(-i_{NL3}g_{m1A} + i_{NL3}g_{m1B}\right)}{\det(j \left(2\omega_1 + \omega_2\right))}$$

$$V_{2,2,1} = \frac{g_m \left(g_L + j \left(2\omega_1 + \omega_2\right) C_L\right) \left(i_{NL3}g_{m1A} - i_{NL3}g_{m1B}\right)}{\det(j \left(2\omega_1 + \omega_2\right))}$$
(5.111)

$$V_{2,2,1} = \frac{g_m \left(g_L + j \left(2\omega_1 + \omega_2\right) C_L\right) \left(i_{NL3g_{m1A}} - i_{NL3g_{m1B}}\right)}{\det(j \left(2\omega_1 + \omega_2\right))}$$
(5.111)

Since  $i_{NL3g_{m1B}} = i_{NL3g_{m1A}}$  we find that

$$V_{1,2,1} = V_{2,2,1} (5.112)$$

Since the output voltage is the difference of the voltage at node 1 and 2 we find

$$V_{out,2,1} = 0 (5.113)$$

This result could have been predicted: the differential output voltage of a differential circuit in which the corresponding components match, does not contain any even-order harmonics of the differential input signal. In other words, frequency components at the frequency  $\pm m\omega_1 \pm n\omega_2$ with m even, do not appear at the output.

Let us now assume that the baseband signal is proportional to the current of the current source at the common emitter and the high-frequency local oscillator signal is applied to the base of the transistors  $Q_{1A}$  and  $Q_{1B}$ . Then we interchange the role of  $\omega_1$  and  $\omega_2$ :  $\omega_1$  is now the local oscillator frequency and  $\omega_2$  is the baseband frequency. If we now want to compute the third-order intermodulation product at  $\omega_1 + 2\omega_2$  then we have to apply other nonlinear current sources as in the previous computations. The phasor of the third-order intermodulation product at the output can be found in a similar way as  $V_{out,2,1}$  was found from equations (5.110) and (5.111):

$$V_{out,1,2} = V_{2,1,2} - V_{1,2,1} = \frac{g_m \left(g_L + (j\omega_1 + 2j\omega_2)C_L\right) \left(2i_{NL3}g_{m1A} - 2i_{NL3}g_{m1B}\right)}{\det(j\omega_1 + 2j\omega_2)}$$
(5.114)

The expression for each nonlinear current source can be derived from Table 5.6 by interchanging the role of  $\omega_1$  and  $\omega_2$ . Hence we find

$$i_{NL3g_{m1A}} = K_{2g_m} V_{43,0,1} V_{43,1,1} + K_{2g_m} V_{43,1,0} V_{43,0,2} + \frac{3}{4} K_{3g_m} V_{43,0,1}^2 V_{43,1,0}$$
 (5.115)

and

$$i_{NL3g_{m1B}} = K_{2g_m} V_{53,0,1} V_{53,1,1} + K_{2g_m} V_{53,1,0} V_{53,0,2} + \frac{3}{4} K_{3g_m} V_{53,0,1}^2 V_{53,1,0}$$
 (5.116)

Most phasors in these equations have already been computed before:  $V_{43,1,0}$  and  $V_{53,1,0}$  are given by equations (5.70) and (5.71), respectively. Expressions for  $V_{43,0,1}$  and  $V_{53,0,1}$  are given in equation (5.79), while equations (5.99) and (5.106) reveal that both  $V_{43,1,1}$  and  $V_{53,1,1}$  are zero. The only phasors that have not been computed yet are  $V_{43,0,2}$  and  $V_{53,0,2}$ . Nodes 4 and 5 are fixed to  $+v_{in}/2$  and  $-v_{in}/2$ , respectively. Hence

$$V_{43,0,2} = V_{53,0,2} = -V_{3,0,2} (5.117)$$

The phasor  $V_{3,0,2}$  is found in a similar way as  $V_{3,2,0}$  was found in equation (5.90):

$$V_{3,0,2} = \frac{(g_L + 2j\omega_2 C_L)^2 \left(i_{NL2g_{m1A}} + i_{NL2g_{m1B}}\right)}{\det(2j\omega_2)}$$
(5.118)

The nonlinear current sources in this equation now correspond to the calculation of second harmonics of  $\omega_2$ . These values are found from the rightmost column of Table 5.5 where the role of  $\omega_1$  and  $\omega_2$  has been interchanged. Hence, the values of  $i_{NL2g_{m1A}}$  and  $i_{NL2g_{m1B}}$  that have to be used in equation (5.118) become

$$i_{NL2g_{m1A}} = \frac{1}{2} K_{2g_m} V_{43,0,1}^2 = \frac{1}{8} \frac{K_{2g_m}}{g_m^2} I_{in}^2$$
 (5.119)

and

$$i_{NL2g_{m1B}} = \frac{1}{2} K_{2g_m} V_{53,0,1}^2 = \frac{1}{8} \frac{K_{2g_m}}{g_m^2} I_{in}^2$$
 (5.120)

Using these values,  $V_{3,0,2}$  from equation (5.118) becomes

$$V_{3,0,2} = \frac{1}{8} \frac{K_{2g_m}}{g_m^3} I_{in}^2 \tag{5.121}$$

With this equation we find for the nonlinear current source  $i_{NL3g_{m1A}}$  in equation (5.115)

$$i_{NL3g_{m1A}} = -\frac{1}{16} \frac{K_{2g_m}^2}{g_m^3} V_{in} I_{in}^2 + \frac{3}{32} \frac{K_{3g_m}}{g_m^2} V_{in} I_{in}^2$$
 (5.122)

and for the nonlinear current source  $i_{NL3g_{m1B}}$  in equation (5.116)

$$i_{NL3g_{m1B}} = \frac{1}{16} \frac{K_{2g_m}^2}{g_m^3} V_{in} I_{in}^2 - \frac{3}{32} \frac{K_{3g_m}}{g_m^2} V_{in} I_{in}^2$$
 (5.123)

which is seen to be the opposite of  $i_{NL3g_{m1A}}$ . Substituting the values of the nonlinear current sources in equation (5.114) yields the final intermodulation product:

$$V_{out,1,2} = \frac{1}{8} \frac{1}{(g_L + (j\omega_1 + 2j\omega_2)C_L)} \left[ -\frac{K_{2g_m}^2}{g_m^3} + \frac{3}{2} \frac{K_{3g_m}}{g_m^2} \right] V_{in} I_{in}^2$$
 (5.124)

With the simple exponential model for the transistor's collector current,  $K_{2g_m}=g_m/(2V_t)$  and  $K_{3g_m}=g_m/(6V_t^2)$ . Hence we obtain

$$V_{out,1,2} = \frac{1}{8} \frac{1}{(g_L + (j\omega_1 + 2j\omega_2)C_L)} \left[ -\frac{g_m^2}{4V_t^2 g_m^3} + \frac{3}{2} \cdot \frac{g_m}{6V_t^2 g_m^2} \right] V_{in} I_{in}^2 = 0$$
 (5.125)

Again it is seen that there is no third-order intermodulation product at the output at the frequency  $\omega_1 + 2\omega_2$ , at least if the simple exponential model for the collector current holds. This can be explained as follows. If in the circuit of Figure 5.6 the current of the current source at the common emitter is kept constant, then a closed-form expression for the output voltage at low frequencies can be derived [Gray 93, Sans 72] if the transistor satisfies the simple exponential relationship and only the collector current of the transistor is taken into account. The output voltage in the time domain is found to be

$$v_{OUT}(t) = -I_{IN}R_L \tanh\left(-\frac{v_{IN}(t)}{2V_t}\right)$$
 (5.126)

It is seen that the output voltage is linearly dependent on the value  $I_{IN}$  of the current source. This is also true if  $I_{IN}$  is a sinusoidal signal. Hence, the output signal does not contain harmonics of this sinusoidal signal.

The reader could question the usefulness of the complicated calculations above that lead to a conclusion that can be formulated directly by looking at the closed-form expression of the input-output relationship, the so-called DC transfer characteristic. First, the example circuit that has been studied here is very simple such that it is still possible to obtain a closed-form

expression for the input-output relationship. For other circuits it might be impossible to obtain such closed-form expression. Also, if in the circuit of Figure 5.6 the base resistance and the output resistance of the transistor are taken into account, then no closed-form expression can be obtained anymore. Next, equation (5.124) formulates the intermodulation product in terms of nonlinearity coefficients. In this equation the value of these coefficients, which of course depend on the model, is not important yet. This means that equation (5.124) is valid even for a complicated model of the collector current. Another argument in favor of the calculation method described here, is that frequency can easily be taken into account. On the other hand, a closed-form expression for the input-output relationship of a circuit including capacitive effects, can in general not be obtained. Finally, the closed-form expression of equation (5.126) is obtained with the assumption of a perfect matching between the transistors  $Q_{1A}$  and  $Q_{1B}$ . If mismatches are present, then a DC transfer characteristic becomes complex or again cannot be generated anymore. With the calculation method described here the inclusion of mismatches does not cause extra problems.

The third-order intermodulation product at  $2\omega_2 + \omega_1$  given in equation (5.124) is not zero if the simple exponential relationship for the collector current does not hold. This occurs at high injection. Nonlinearity coefficients for the high-injection operating region are discussed in Chapter 6. However, a bipolar transistor is seldom biased in the high-injection region. As a result, third-order intermodulation distortion will still be small, even with a more exact model for the collector current.

The third-order intermodulation product given in equation (5.124) can also be used for simple MOS mixer as shown in Figure 5.10. If the drain current of the MOS transistor satisfies

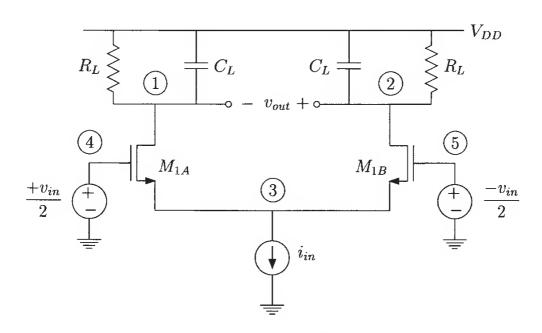


Figure 5.10: A simple MOS mixer.

the square law, then we find from Table 3.2 that  $g_m=\beta(V_{GS}-V_T)$ ,  $K_{2g_m}=\beta/2$  and  $K_{3g_m}=0$ .

The third-order intermodulation product now becomes

$$V_{out,1,2} = -\frac{1}{32} \frac{1}{(g_L + (j\omega_1 + 2j\omega_2)C_L)} \frac{V_{in} \cdot I_{in}^2}{\beta (V_{GS} - V_T)^3}$$
(5.127)

Consider now the expression of the wanted signal, which, in the case of an upconverter is the component at the sum frequency  $\omega_1 + \omega_2$ . Equation (5.93) obtained for the bipolar transistor mixer can also be used here. Using the appropriate values of the nonlinearity coefficients we obtain

$$V_{out,1,1} = \frac{1}{4(V_{GS} - V_T)} \frac{1}{(g_L + j(\omega_1 + \omega_2) C_L)} V_{in} I_{in}$$
 (5.128)

The ratio of the third-order intermodulation product and the conversion gain is then found to be

$$\frac{V_{out,1,2}}{V_{out,1,1}} = -\frac{1}{8} \frac{1}{\beta (V_{GS} - V_T)^2} \frac{g_L + (j\omega_1 + j\omega_2)C_L}{g_L + (j\omega_1 + 2j\omega_2)C_L} I_{in}$$
(5.129)

When this ratio is set equal to one and solved for  $I_{in}$  then the third-order intercept point is found:

$$IP_{3} = 8\beta(V_{GS} - V_{T})^{2} \cdot \left| \frac{g_{L} + (j\omega_{1} + 2j\omega_{2})C_{L}}{g_{L} + (j\omega_{1} + j\omega_{2})C_{L}} \right|$$
 (5.130)

This value has the dimensions of a current. In most cases, the baseband signal is a voltage that is first converted to a current by a transconductor. This conversion will of course cause nonlinear distortion that has not been taken into account in this simple example.

#### 5.3.4 Interpretation and factorization

Up till now we have seen two methods for the computation of nonlinear responses: in this section we considered the direct computation of nonlinear responses. In Section 5.2 we computed Volterra kernels from which the nonlinear responses are derived. It is seen that the two approaches are very similar. Hence the results can be interpreted in a similar way as described in Section 5.2.7. Furthermore, the denominators of the nonlinear responses can be factorized as well, in the same way as described in Section 5.2.8. As a result, the numerator of any harmonic can be computed separately from the denominator.

Table 5.8 lists the numerators of the nonlinear current sources of order two that have to be applied in order to compute the numerator of any second harmonic of  $\omega_1$  in the circuit. Hereby the numerator of the phasor  $V_{i,k,m}$  is represented as  $Vn_{i,k,m}$ . The denominator for all second harmonics of  $\omega_1$  in the circuit is given by

denominator of 2nd harmonic of 
$$\omega_1 = 2 \left( \det(j\omega_1) \right)^2 \det(2j\omega_1)$$
 (5.131)

and for the second harmonic distortion

denominator of 
$$HD_2 = 2 \det(j\omega_1) \det(2j\omega_1)$$
 (numerator 1st-order response) (5.132)

	numerator of	
type of nonlinearity	nonlinear current source	
	for response at $2\omega_1$	
(trans)conductance	$K_{2g_1} (Vn_{i,1,0})^2$	
capacitor	$2j\omega_1 K_{2_{C_1}} (Vn_{i,1,0})^2$	
two-dimensional	-	
conductance	$K_{2g_1\&g_2} V n_{i,1,0} V n_{j,1,0}$	
(only cross-terms)		

Table 5.8: Numerators of the nonlinear second-order current sources for the basic nonlinearities to directly compute the numerator of second harmonics at  $2\omega_1$ . The controlling voltages are  $v_i$  for the nonlinear (trans)conductance and the nonlinear capacitor and  $v_i$  and  $v_j$  for the two-dimensional conductance.

For the computation of third harmonics in a circuit the numerators of the nonlinear current sources of order three that have to be applied are listed in Table 5.9. The common denominator for all third harmonics of  $\omega_2$  is given by

denominator of 3rd harmonic of 
$$\omega_2 = 4 \left( \det(j\omega_2) \right)^3 \det(2j\omega_2) \det(3j\omega_2)$$
 (5.133)

and for the third harmonic distortion

denominator of  $HD_3 = 4 \left( \det(j\omega_2) \right)^2 \det(2j\omega_2) \det(3j\omega_2)$  (numerator of 1st-order response) (5.134)

# 5.4 Symbolic computation of harmonics and intermodulation products

In the previous sections we have discussed the Volterra series method and the direct calculation method of nonlinear responses. In both cases, the results are obtained by repeatedly solving a linear network. Hence, it is possible to obtain closed-form expressions for the nonlinear responses in terms of one or more frequency variables, the small-signal parameters and the nonlinearity coefficients.

type of nonlinearity	numerator of	
tjpo oz	nonlinear current source for response at $3\omega_2$	
(trans)conductance	$2K_{2g_1}Vn_{i,0,1}Vn_{i,0,2} + \det(2j\omega_2)K_{3g_1}Vn_{i,0,1}^3$	
capacitor	$3j\omega_2 \left[ 2K_{2C_1}Vn_{i,0,1}Vn_{i,0,2} + \det(2j\omega_2)K_{3C_1}Vn_{i,0,1}^3 \right]$	
two-dimensional	$K_{2_{g_1}\&g_2} \bigg[ Vn_{i,0,1}Vn_{j,0,2} + Vn_{i,0,2}Vn_{j,0,1} \bigg]$	
conductance		
(only cross-terms)	$+ \det(2j\omega_2) K_{3_{2g_1\&g_2}} V n_{i,0,1}^2 V n_{j,0,1} + \det(2j\omega_2) K_{3_{g_1\&2g_2}} V n_{i,0,1} V n_{j,0,1}^2$	
three-dimensional		
conductance	$\det(2j\omega_2) K_{3_{g_1}\&g_2\&g_3} V n_{i,0,1} V n_{j,0,1} V n_{k,0,1}$	
(only cross-terms)		

Table 5.9: Numerators of the nonlinear third-order current sources for the basic nonlinearities to compute the numerator of the third harmonic at  $3\omega_2$ . The controlling voltages are  $v_i$  for the nonlinear (trans)conductance and the nonlinear capacitor,  $v_i$  and  $v_j$  for the two-dimensional conductance and  $v_i$ ,  $v_j$  and  $v_k$  for the three-dimensional conductance.

A symbolic expression that describes the behavior of a circuit gives information that is complementary to information supplied by numerical simulations. Whereas the latter information can be plotted on a graph that can be interpreted, a simple closed-form expression can immediately show the dominant circuit parameters. Hereby an exact expression is seldom required since this is very often too lengthy or too complicated: a circuit designer is often willing to pay the price of a limited accuracy in order to get an approximate but interpretable expression.

Section 4.8.6 and the examples of Section 5.2 and 5.3 already contain some hand calculations that result in closed-form expressions for nonlinear responses. Although these examples are conceptually very simple, the calculations are already quite involved. For larger circuits of practical interest, containing several nonlinearities, hand calculations become tedious and error-prone. This situation can be relieved if the calculations can be automated: in this way, a circuit designer can concentrate on the interpretation of the expression rather than on the generation of the expression.

Such automation of calculations can be obtained with a symbolic network analysis program.

Such program reads in a netlist of an analog circuit and generates a closed-form expression for the characteristic that the user wants to simulate. Many symbolic network analysis programs have already been reported for linear or linearized circuits [Giel 89, Fern 91b, Man 91, Wier 89, Kon 88, Hass 89, Hue 89, Somm 93, Neb 95, Wamb 95, Yu 96]. This approach is feasible for linear or linearized circuits. For nonlinear circuits this is not possible since in general closed-form expressions cannot be obtained for nonlinear behavior. However, if we restrict the analysis to weakly nonlinear behavior, then the previous sections have shown that closed-form expressions can be obtained.

#### 5.4.1 Symbolic network analysis of linearized analog circuits

Just as we did in the calculations of Sections 5.2 and 5.3 symbolic expressions for a transfer function are computed with the rule of Cramer. In this way, a transfer function is found as a ratio of two determinants. Classically, a determinant is computed by developing it along a row or a column. Many researchers have developed alternative methods [Mas 53, May 57, Coat 58, Chen 65, Ald 73, Sann 80] for this approach. For example, in [May 57] the different terms of a determinant are computed by enumerating spanning trees that are common to two graphs that correspond to the circuit under consideration. Clearly, many of these methods are complicated for hand calculations but they turn out to be very efficient in computer programs for symbolic network analysis. For example, the method that enumerates common spanning trees is widely used in modern symbolic network analyzers [Wamb 95, Yu 96]. The interested reader is referred to specialized literature [Giel 91, Giel 94a, Wamb 96, Wamb 97].

The formula for a transfer function that is generated by a symbolic network analysis program formally looks as follows:

$$T = \frac{f_0(\mathbf{x}) + sf_1(\mathbf{x}) + s^2 f_2(\mathbf{x}) + \ldots + s^n f_n(\mathbf{x})}{g_0(\mathbf{x}) + sg_1(\mathbf{x}) + s^2 g_2(\mathbf{x}) + \ldots + s^m g_m(\mathbf{x})}$$
(5.135),

in which  $\mathbf{x}^T = \{x_1, x_2, ... x_Q\}$  is the vector of symbolic circuit elements and the  $f_i$  (i = 0, ..., n) and  $g_j$  (j = 0, ..., m) are sums of products of symbolic circuit elements. In some applications it is useful to compute a transfer function as a function of s only, while the circuit parameters are treated as numbers. In this case, the coefficients  $f_i$  and  $g_j$  in equation (5.135) are numbers. For such applications dedicated methods exist, such as the polynomial interpolation method explained in [Vlach 83]. This polynomial interpolation method is also used in Section 5.4.2.1 to eliminate insignificant nonlinearities. The result of the polynomial interpolation method is seldom of direct interest to a circuit designer: usually, a circuit designer is more interested in fully symbolic expressions.

With the advent of powerful computers and due to many research efforts in the domain of CAD for analog integrated circuits, symbolic network analysis received a renewed interest in the late eighties. It was (and it is still) believed that symbolic network analysis programs are able to generate closed-form expressions that can be helpful during the design of analog integrated circuits: either the expressions can be interpreted or they can be used in design automation applications where the expressions are evaluated repeatedly [Giel 91]. In both cases, the goal is to accelerate the design of a given analog circuit.

Symbolic network analysis programs are faced to two major problems. A first problem is that the number of terms of the exact transfer function increases exponentially with the size (number of nodes and number of circuit elements) of the circuit. For example, in [Wamb 97] it is shown that the exact symbolic expression of the differential-mode gain for the simple operational amplifier of Figure 5.11 contains about 700 terms. This symbolic expression contains the fre-

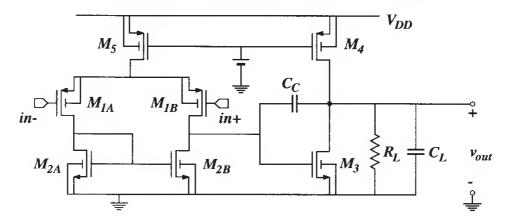


Figure 5.11: A CMOS Miller-compensated operational amplifier.

quency variable s and the elements of the small-signal equivalent circuit of every transistor. The symbolic expression of the differential-mode gain for the fully-differential BiCMOS operational amplifier of Figure 5.12 contains about  $10^{12}$  terms [Wamb 95]. Clearly, these terms cannot all be generated, only the number of terms can be estimated as explained in [Wamb 97]. Another

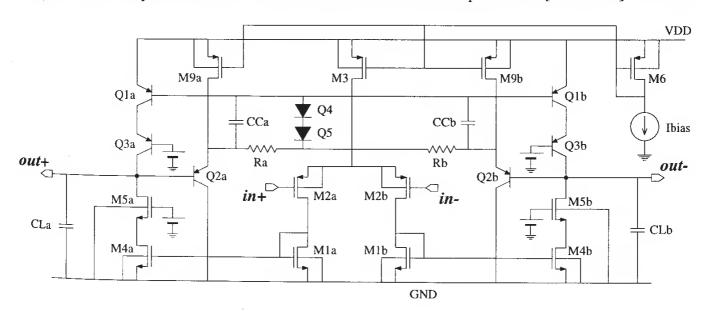


Figure 5.12: A fully-differential BiCMOS operational amplifier.

problem is the generation of an interpretable expression. Even for a small circuit the number of terms in a symbolic transfer function is too large to allow an easy interpretation.

In order to tackle these problems, many symbolic analysis programs have been developed from the late eighties on that were able to generate approximate symbolic expressions [Giel 89, Fern 91b, Man 91, Wier 89, Kon 88, Hass 89, Hue 89, Somm 93, Neb 95, Wamb 95, Yu 96]. The approximation is made by making use of numerical values for the different circuit elements. These numerical values can originate from estimations or from a sized circuit. The approximation proceeds by pruning the insignificant terms from the exact symbolic expression for a transfer function. For such approximation, three major strategies have been reported in literature. A first strategy, that has been used in the first symbolic simulators that supported approximation [Giel 89, Fern 91b], is the so-called *simplification after generation* strategy. With this strategy the symbolic transfer function is first computed exactly. Afterwards, the insignificant terms are pruned. This strategy only works for small circuits containing no more than ten nodes. For large circuits the number of terms is too large such that a complete generation of the exact expression is not feasible.

A better strategy for large circuits is the so-called *simplification before generation* strategy [Somm 93, Yu 96, Hsu 94]. With this strategy the circuit netlist is simplified before the actual symbolic computations begin. The simplification is based on numerical computations in which the circuit elements have a numerical value. For example, a simplification before generation strategy can eliminate the network elements of the bias circuitry that is outside the signal path. Indeed, since the transfer function of interest is nearly independent of these network elements, it is a good idea to remove these elements. As another example, a simplification before generation approach can eliminate capacitors at low-impedance nodes when one is interested in a simulation at low frequencies. The result of this simplification is a circuit with less elements. This circuit can be analyzed symbolically in a much easier way than the original circuit since it is smaller.

Another simplification strategy that has caused a major breakthrough in symbolic analysis of analog integrated circuits of practical size is the *simplification during generation* strategy [Wamb 92, Wamb 94a, Wamb 95, Yu 96]. With this approach only the dominant terms of a transfer function are generated. The terms are generated in decreasing order of magnitude. The generation continues until the sum of generated terms is sufficiently close to the exact numerical result.

In order to get a flavor of symbolic network analysis and approximation, an example is presented in the next section.

**Example:** differential-mode gain of a fully-differential operational amplifier Figure 5.13 and 5.14 shows the dominant terms of the numerator and the denominator of the low-frequency differential-mode gain of the fully-differential BiCMOS operational amplifier of Figure 5.12. These expressions have been generated with a simplification during generation approach [Wamb 9]. The exact expression contains about 10<sup>12</sup> terms and hence it is impossible to generate this.

In these two expressions, the terms are sorted in decreasing order. The terms contain conductances, for example  $G_a=1/R_a$  and transconductances. Matching elements have been represented by the same symbol. This means for example that the transconductance of transistors  $M_{2A}$  and  $M_{2B}$  are both represented by the same symbol  $g_{m_{M1}}$ . The element  $g_{eq2}$  is a conductance that

$$\begin{array}{l} 4\,g_{m_{Q2}}^2\,g_{m_{Q3}}^2\,g_{m_{M6}}\,g_{m_{M2}}^2\,g_{m_{M5}}^2\,g_{m_{M1}}\,g_{m_{Q5}}\,g_{m_{M4}}\,g_{m_{Q4}}\,G_a\,g_{m_{Q1}}\\ \\ +\ 8\,g_{m_{Q2}}^2\,g_{m_{Q3}}^2\,g_{m_{Q5}}\,g_{m_{M2}}^2\,g_{m_{b_{M5}}}\,g_{m_{M5}}\,g_{m_{M1}}\,g_{m_{M6}}\,g_{m_{M4}}\,g_{m_{Q4}}\,G_a\,g_{m_{Q1}}\\ \\ +\ 4\,g_{m_{Q1}}\,g_{m_{Q3}}^2\,g_{m_{Q5}}\,g_{m_{M2}}^2\,g_{m_{Q2}}^2\,g_{m_{b_{M5}}}^2\,g_{m_{M1}}\,g_{m_{M6}}\,g_{m_{M4}}\,g_{m_{Q4}}\,G_a\\ \\ +\ 4\,g_{m_{Q2}}^2\,g_{m_{Q3}}^2\,g_{o_{M1}}\,g_{m_{M2}}^2\,g_{m_{Q5}}^2\,g_{m_{M5}}^2\,g_{m_{M6}}\,g_{m_{M4}}\,g_{m_{Q4}}\,G_a\,g_{m_{Q1}} \end{array}$$

Figure 5.13: Approximate expression for the numerator of the low-frequency gain of the circuit of Figure 5.12.

$$\begin{array}{l} 8g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{o_{M1}} \ g_{m_{M2}} \ g_{m_{M1}} \ g_{o_{M4}} \ g_{m_{Q5}} \ G_a \ g_{m_{M5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ G_a \ g_{m_{M1}}^2 \ g_{o_{M4}} \ g_{m_{Q5}} \ g_{o_{M5}}^2 \ g_{m_{M6}} \ g_{m_{M2}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 8 \ g_{\pi_{Q2}} \ g_{m_{Q2}} \ g_{m_{Q3}}^2 \ G_a^2 \ g_{m_{M2}} \ g_{m_{M1}}^2 \ g_{m_{M5}} \ g_{m_{M5}} \ g_{m_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{\pi_{Q2}} \ g_{m_{Q3}}^2 \ g_{m_{M1}}^2 \ g_{m_{M2}} \ g_{o_{M4}} \ g_{m_{Q5}} \ G_a \ g_{m_{M5}}^2 \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ G_a \ g_{m_{M1}}^2 \ g_{o_{M4}} \ g_{m_{bM5}} \ g_{m_{Q5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{\pi_{Q2}}^2 \ g_{m_{Q3}}^2 \ G_a^2 \ g_{m_{M2}} \ g_{m_{M1}}^2 \ g_{m_{M5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ G_a^2 \ g_{m_{M2}} \ g_{m_{M1}}^2 \ g_{m_{Q5}} \ g_{m_{M6}}^2 \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{m_{M2}}^2 \ g_{m_{M1}}^2 \ g_{o_{M4}}^2 \ g_{m_{Q5}} \ G_a \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{m_{M3}}^2 \ g_{m_{M1}}^2 \ g_{o_{M4}}^2 \ g_{m_{Q5}} \ G_a \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{m_{M3}}^2 \ g_{m_{M1}}^2 \ g_{o_{M4}}^2 \ g_{m_{M5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{m_{M1}}^2 \ g_{o_{M4}}^2 \ g_{m_{M5}} \ g_{m_{M6}} \ g_{o_{M5}} \ g_{m_{M6}} \ g_{m_{Q4}} \ g_{m_{Q1}} \\ + \ 4 \ g_{m_{Q2}}^2 \ g_{m_{Q3}}^2 \ g_{m_{M1}}^2 \ g_{o_{M4}}^2 \ g_{m_{M6}}^2 \ g_{m_{M$$

Figure 5.14: Approximate expression for the denominator of the low-frequency gain of the circuit of Figure 5.12.

represents the parallel connection of two output conductances of transistors:

$$g_{eq2} = g_{o_{M9}} + g_{o_{Q2}} (5.136)$$

Here  $g_{o_{M9}}$  is the output conductance of transistor  $M_{9a}$  or  $M_{9b}$  and  $g_{o_{Q2}}$  is the output of transistor  $Q_{2A}$  or  $Q_{2B}$ .

It is seen that the expression of the numerator and denominator, although they contain just a few terms, are still too complicated to interpret. In order to improve the interpretability, extra postprocessing is required. This postprocessing can be part of a symbolic simulator as well.

As a first postprocessing step, extra terms are removed. For a specified error this is possible here since the error on the ratio of the numerator and denominator is much smaller than the error on numerator and denominator individually. In this example, the error on numerator and denominator individually is about 20% and the error on the ratio about 1%. After the removal of extra terms, we keep the two largest terms from Figure 5.13 for the numerator, together with the five largest terms from Figure 5.14 for the denominator. The error on the resulting ratio is only two percent.

From the above symbolic expressions it is seen that every product term in the numerator and denominator consists of fifteen symbols. These product terms contain elements from the bias circuitry (like  $g_{m_{M6}}$ ) or from the common-mode feedback circuit (like  $g_{m_{Q4}}$  and  $g_{m_{Q5}}$ ) that do not have much influence on the differential-mode gain. After factorization, however, they disappear. The final result, after post-processing is now:

$$\frac{g_{m_{Q2}}g_{m_{M2}}g_{m_{M4}}\left(g_{m_{M5}}+g_{mb_{M5}}\right)}{g_{m_{M1}}\left(g_{m_{Q2}}g_{o_{M4}}g_{o_{M5}}+g_{\pi_{Q2}}(G_a+g_{eq2})(g_{m_{M5}}+g_{mb_{M5}})\right)}$$
(5.137)

This can be rewritten as

$$\frac{g_{m_{M2}}}{g_{m_{M1}}} \cdot \frac{g_{m_{M4}}}{\frac{g_{o_{M4}} g_{o_{M5}}}{g_{m_{M5}} + g_{mb_{M5}}} + \frac{G_a + g_{eq2}}{\beta_{Q2}}}$$
(5.138)

The interpretation is now as follows: the input transistors  $M_{2A}$  and  $M_{2B}$  are each loaded with a diode-connected transistor, yielding a gain of  $g_{m_{M2}}/g_{m_{M1}}$ . The gain of the second stage is  $g_{m_{M4}}/g_{LOAD}$  with  $g_{LOAD}$  being the total conductance seen at the output node in differential mode. This is the sum of the output conductance of the MOS cascode stage  $M_{4a}-M_{5a}$  or  $M_{4b}-M_{5b}$  and the input conductance of the emitter follower. The output conductance of the bipolar cascode is too small and does not contribute significantly to  $g_{LOAD}$ .

#### 5.4.2 Symbolic analysis of weakly nonlinear analog circuits with ISAAC

The calculation method of Section 5.3 has been built into the symbolic simulator ISAAC [Giel 89, Giel 91]. This simulator can compute approximate symbolic expressions for the AC characteristics of analog circuits. The approximation is performed with a simplification *after* generation strategy. As a result, the simulator can only be used for fairly small circuits, having at most tento twelve transistors, depending on the circuit topology. This is not a limitation for the examples treated in this book, since these are rather simple.

The nonlinear responses are computed with ISAAC by treating the numerator and the denominator of the responses separately. This is possible, as explained in Sections 5.2.8 and 5.3.4. The denominator of any nonlinear response of order two or three is a product of determinants that are nothing else but the determinant of the admittance matrix (or (C)MNA matrix) of the linearized network, but evaluated at other frequencies. The numerator of a harmonic or intermodulation product can be computed by combining the numerators of several transfer functions, either from the input to a controlling voltage or from a nonlinear current source to the output or a controlling voltage. The final result is a nested expression, which can be expanded afterwards.

The numerators of the involved transfer functions can be calculated with a routine to compute determinants of matrices with symbolic entries. In ISAAC determinants are computed by a development along rows or columns. However, huge expressions are generated in this way since a practical circuit contains a lot of basic nonlinearities and each nonlinearity gives rise to a nonlinear current source, whose expression can already be quite complicated.

In order to manage this complexity, two simplification procedures are used. In a first simplification *before* generation step, which is described in Section 5.4.2.1, numerical computations are used to eliminate the nonlinearities that are unimportant in the frequency range of interest. After this elimination, nonlinearity coefficients that do not play a role are set equal to zero.

After the first simplification step, a symbolic expression is computed. This can be performed either with a simplification *after* generation procedure or with a simplification *during* generation procedure. In the program ISAAC a simplification *after* generation strategy is used [Giel 89, Giel 91].

#### 5.4.2.1 Elimination of unimportant nonlinearities

The contribution of every nonlinearity coefficient to the nonlinear response of interest is first calculated as a function of frequency while circuit elements are represented by their numerical values. This can be accomplished very efficiently with the polynomial interpolation method, described in [Vlach 83]. The resulting contribution consists in this way of a product of polynomials in the frequency variable s. The coefficients of these polynomials are numbers. In the previous sections it was seen that the numerator of a network function can be reused for computations at different orders. For example, the numerator of the transfer function from a nonlinear current source to the output is used for all orders higher than one, except that the frequency variables are different. Hence, if numerators of network functions are computed as a function of a frequency variable, then this variable is easily adapted appropriately to the order under consideration. In this way, lower-order results are maximally reused.

In addition to the calculation of the contribution of every nonlinearity coefficient, the total response is also calculated as a function of a frequency variable by making the sum of the different contributions (see equation (5.38)). The results, both the individual contributions and the total result are now used to determine which nonlinearities can be eliminated.

First, the frequency interval of interest is discretized. Then for every frequency point the following procedure is executed. The nonlinearities are sorted in decreasing order in two arrays, each according to a different key. The first key is the absolute value of the real part of the contribution of the nonlinearity to the output response at the frequency under consideration. The second key is the absolute value of the imaginary part. Then it is determined how many nonlinearities are required such that the sum of the real (or imaginary) part of their contributions is sufficiently close — up to a user-provided error — to the real or imaginary part of the total response at that frequency. In this way, just enough nonlinearities are included, starting with the ones that have the largest real or imaginary part, until the error is sufficiently small. Next, a following frequency point is examined. At this frequency point the nonlinearities that have been included at any previous frequency point are always included. If the real or imaginary part of any other nonlinearity that has not been previously included, turns out to be larger than the real or imaginary part of any previously included nonlinearity, then that nonlinearity is included as well. Now the sum of the real parts of the included contributions is compared to the real part of the total response. The same is done for the imaginary parts. If the accuracy is not sufficient for either the real or imaginary part, then additional nonlinearities will be included.

After this procedure has been executed for every frequency point, a second turn is performed

over the frequency points: it is necessary that the accuracy is checked again at every frequency point since a nonlinearity that has been included at a higher frequency can increase the deviation between the sum of the real or imaginary parts of the nonlinearities and the real or imaginary part of the total value at that frequency.

The elimination of insignificant nonlinearities as explained above often reduces significantly the number of nonlinearities in practical analog integrated circuits as we will see in the examples of Chapter 8.

Note that the intermediate results obtained in this step, namely a knowledge of the significant nonlinearities, provides information which is already much more valuable than simulation results obtained from SPICE-like simulators with the .DISTO command, which do not select the significant nonlinearities at all. This knowledge, together with a plot of the different contributions and the total response as a function of frequency can already yield enough insight such that the used does not need a symbolic analysis anymore. This will also be illustrated in Chapter 8.

#### 5.4.2.2 Generation of the approximate symbolic subexpressions

Equation (5.38) reveals that the expressions for the weakly nonlinear behavior of the circuit are nested. In order to simplify such expression using numerical values of the symbols, the approximation algorithm for nested expressions, described in [Fern 93a]. More details can be found in [Wamb 90, Wamb 91a, Wamb 91b, Wamb 96].

#### 5.5 Simple example circuits

In this section we consider a few simple networks for which harmonics and/or Volterra kernel transforms are computed. The networks are simpler than the networks that are considered in Chapter 8. The goal of this section is to get insight in the operation of the small nonlinear networks and to further illustrate the two calculation methods that have been discussed in this chapter. The two simple networks are a nonlinear voltage divider and a nonlinear capacitive current divider. For the first network we choose to compute Volterra kernels instead of responses. Since this network is memoryless, the Volterra kernels can be computed as simply as harmonics or intermodulation products. For the second network we directly compute responses.

#### 5.5.1 Nonlinear resistive voltage divider

Figure 5.15 depicts a nonlinear resistive voltage divider. The output of interest is the voltage a node a.

We will describe the resistors in conductance form, which means that the current will be expressed as a function of the voltage over the element. The description of resistor  $R_1$  therefore

$$i_{in} = f_1(v_1) (5.139)$$

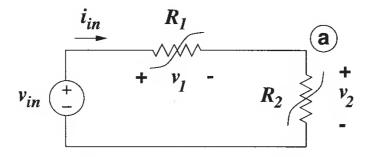


Figure 5.15: A nonlinear resistive voltage divider.

in which  $f_1$  is a nonlinear function. The current  $i_{in}$  and the voltage  $v_1$  are explained in Figure 5.15. Developing the function  $f_1$  into a power series that is broken down after the third term yields

$$i_{in} = g_1 v_1 + K_{2q_1} v_1^2 + K_{3q_1} v_1^3 (5.140)$$

Here the symbol  $g_1$  is the conductance that is the inverse of the AC value of resistance  $R_1$ . The description in conductance form of resistor  $R_2$  is given by

$$i_{in} = f_2(v_2) (5.141)$$

As seen in Figure 5.15 the current through  $R_2$  is the same as the current through  $R_1$ . The power series description of  $R_2$  is given by

$$i_{in} = g_2 v_2 + K_{2q_2} v_2^2 + K_{3q_2} v_2^3 (5.142)$$

**First-order kernels** The first step in the procedure to determine the Volterra kernels is the computation of the linear transfer functions. This is performed with the linearized equivalent of Figure 5.15, which is shown in Figure 5.16.

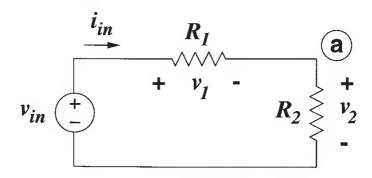


Figure 5.16: Linearized equivalent of the resistive voltage divider.

Applying Kirchoff's current law at node a yields

$$g_1(v_a - v_{in}) + g_2 v_a = 0 (5.143)$$

From this equation one finds

$$v_a = \frac{g_1}{g_1 + g_2} v_{in} \tag{5.144}$$

The value of  $v_a$  obtained in this linear circuit is also the value of the linear transfer function of the voltage  $v_2$  over resistor  $R_2$  if at least  $v_{in}$  is set equal to one. This transfer function, denoted by  $H_{12}$ , is thus given by

$$H_{1_2} = H_{1_a} = \frac{g_1}{g_1 + g_2} \tag{5.145}$$

The first subscript in  $H_{12}$  and  $H_{1a}$  indicates the order of the transfer function, whereas the second subscript corresponds to the numbering of the node voltages.

The linear transfer function for voltage  $v_1$  that controls  $R_1$  is easily found to be

$$H_{1_1} = \frac{g_2}{g_1 + g_2} \tag{5.146}$$

This transfer function will be required for the computation of the nonlinear current source of order two that corresponds to  $R_1$ .

**Second-order kernels** Next, the second-order Volterra kernels are determined in the circuit. To this purpose, the linearized circuit of Figure 5.16 is excited with the nonlinear current sources of order two that correspond to  $R_1$  and  $R_2$ , whereas the external voltage source is neutralized. The resulting circuit is shown in Figure 5.17.

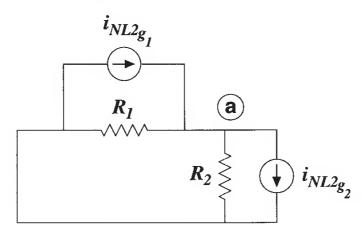


Figure 5.17: Circuit for the computation of the second-order kernels.

The values of the nonlinear current sources of order two can be found by combining the general expression of a nonlinear current source for a nonlinear conductance in Table 5.1, with the expressions for the first-order transfer functions, equations (5.145) and (5.146). For the nonlinear current source of order two  $i_{NL2g_1}$ , that corresponds to  $R_1$ , we find in this way

$$i_{NL2g_1} = K_{2g_1} \left(\frac{g_2}{g_1 + g_2}\right)^2 \tag{5.147}$$

whereas for the source  $i_{NL2q_1}$ , that corresponds to  $R_1$ , we find

$$i_{NL2g_2} = K_{2g_2} \left(\frac{g_1}{g_1 + g_2}\right)^2 \tag{5.148}$$

Applying Kirchoff's current law at node a yields

$$(g_1 + g_2) v_a - i_{NL2g_1} + i_{NL2g_2} = 0 (5.149)$$

The unknown voltage  $v_a$  in this equation is nothing else but the second-order kernel transform of the voltage  $v_a$  or  $v_2$ , denoted by  $H_{2_2}$ . Hence we find from equation (5.149)

$$H_{2_2} = \frac{i_{NL2g_1} - i_{NL2g_2}}{g_1 + g_2} \tag{5.150}$$

Using equations (5.147) and (5.148) we find then for  $H_{22}$ 

$$H_{2_2} = H_{2_a} = \frac{1}{(g_1 + g_2)^3} \left( K_{2g_1} g_2^2 - K_{2g_2} g_1^2 \right)$$
 (5.151)

and for the second harmonic distortion we find

$$HD_2 = \frac{V_{in}}{2} \cdot \frac{H_{2a}}{H_{1a}} = \frac{V_{in}}{2} \cdot \frac{\left(K_{2g_1}g_2^2 - K_{2g_2}g_1^2\right)}{\left(g_1 + g_2\right)^2 \cdot g_1} \tag{5.152}$$

It is seen that the contributions of the two nonlinearities are opposite. It is interesting to check whether these contributions can cancel. Assume that the two resistors are tracking nonlinearities. According to equations (3.81) and (3.82) this means

$$g_2 = ag_1 (5.153)$$

$$K_{2g_2} = aK_{2g_1} (5.154)$$

where a is a constant. In this case,  $H_{2a}$  reduces to

$$H_{2a} = \frac{a(a-1)}{(1+a)^3} \frac{K_{2g_1}}{g_1}$$
 (5.155)

and this is zero only if a=1. In other words, the second-order kernel transform and hence the second-order responses at the output of a nonlinear resistive voltage divider with two tracking nonlinear resistors are zero only if the resistors are identical.

Equation (5.151) is now evaluated for a practical example. Assume that resistor  $R_2$  is the resistor  $r_{\pi}$  of a bipolar transistor and  $r_B$  is the nonlinear base resistance of this transistor. Detailed descriptions of these nonlinearities will be given in Chapter 6. For this example, we will use the simple expressions for the nonlinearity coefficients that describe  $g_{\pi} = 1/r_{\pi}$  from equations (3.20) and (3.22). For coefficients that describe the base resistance realistic values will be used that will be justified later in Section 6.4.

For a collector current of 1mA and a transistor beta  $\beta_F$  of 100 we find at room temperature

$$g_{\pi} = \frac{I_C}{\beta_E V_t} = 4 \times 10^{-4} A/V \tag{5.156}$$

$$K_{2g_{\pi}} = \frac{g_{\pi}}{2V_t} = 8 \times 10^{-3} A/V^2 \tag{5.157}$$

A value of  $4\times10^{-4}A/V$  for  $g_{\pi}$  corresponds to a value of  $2500\Omega$  for  $r_{\pi}$ .

For the AC conductance  $g_B$  that corresponds to the base resistance we take a value of  $5\times 10^{-3}A/V$ , corresponding to an AC resistance of  $200\Omega$ . At fairly high bias currents the base resistance starts to fall off with the current due to several effects. Hence the base resistance becomes nonlinear. This will be explained in Section 6.4. Circuit designers might argue that this bias region is not very realistic: in practice one would take a larger bipolar transistor for the given bias current, such that the base resistance is smaller and that is does not fall off yet with the current. Nevertheless, we assume a nonlinear base resistance here just for illustration purposes. A value of  $2A/V^2$  is taken for  $K_{2g_B}$ . From the above values we find for the normalized nonlinearity coefficients of  $g_{\pi}$  and  $g_B$ 

$$K_{2q_{\pi}}' = 20V^{-1} \tag{5.158}$$

$$K_{2q_B}' = 400V^{-1} (5.159)$$

This means that  $g_B$  is "more nonlinear" than  $g_\pi$  for the given numerical values.

With the above values  $H_{2a}$  evaluates to  $0.762V^{-1}$  and  $HD_2$  for an input amplitude of 1V is 0.411. If the base resistance would be linear, such that  $K_{2g_B}=0$ , then  $H_{2a}$  is equal to  $-1.27V^{-1}$ . This value is larger than with a nonlinear base resistance and it has an opposite sign. This again illustrates the compensating effects of the two nonlinearities.

The ratio of the two contributions is found from equation (5.151) to be

$$\left| \frac{\text{contribution of } K_{2g_B}}{\text{contribution of } K_{2g_{\pi}}} \right| = \frac{K_{2g_B} g_{\pi}^2}{K_{2g_{\pi}} g_B^2}$$
 (5.160)

In our numerical example this ratio is 1.6. In reality, the base resistance consists of an intrinsic part  $r_{Bi}$  and an extrinsic part  $r_{Bex}$  which can be considered as linear. This linear part has been neglected thus far. However, it will decrease the contribution of  $K_{2g_{Bi}}$ , which describes the second-order nonlinearity of  $r_{Bi}$ , compared to the contribution of  $K_{2g_{\pi}}$ . This has been checked with a symbolic computation using ISAAC of the second-order kernel of the voltage at node a in the circuit of Figure 5.18. With ISAAC the ratio of the contribution of  $K_{2g_{Bi}}$  and the contribution of  $K_{2g_{\pi}}$  to the second-order kernel of the voltage at node a is found to be

$$\left| \frac{\text{contribution of } K_{2g_{Bi}}}{\text{contribution of } K_{2g_{\pi}}} \right| = \frac{K_{2g_{Bi}}g_{\pi}^{2}}{K_{2g_{\pi}}g_{Bi}^{2}} \frac{g_{Bex}}{g_{Bex} + g_{Bi}}$$
(5.161)

In this equation  $g_{Bex}=1/r_{Bex}$ . If we take for  $g_{Bi}$  and  $K_{2g_{Bi}}$  the same values as for  $g_{B}$  and  $K_{2g_{B}}$ , namely  $5\times10^{-3}A/V$  and  $2A/V^2$ , and  $r_{Bex}$  is taken equal to  $150\Omega$ , then the ratio in equation (5.161) is now reduced to 0.89 instead of 1.6 with equation (5.160).

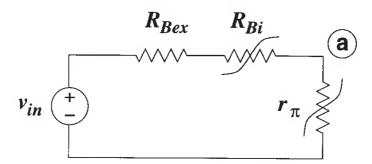


Figure 5.18: Voltage division between the (linear) extrinsic part  $r_{Bex}$  of the base resistance of a bipolar transistor, its nonlinear intrinsic part  $r_{Bi}$  and the nonlinear  $r_{\pi}$ .

From equation (5.161) it is seen that when  $g_{Bex}$  goes to zero, which corresponds to an excitation of the voltage divider by a current source, then the nonlinearity of  $r_{Bi}$  does not contribute anymore: only  $r_{\pi}$  gives a contribution in this case.

Third-order kernels For the computation of the third-order kernel transform of the voltage at node a, we will need the second-order kernel of all voltages that control a nonlinearity, since the nonlinear current sources of order three depend on these second-order kernels. Hence, in addition to  $H_{2_2}$ , we also need the second-order kernel of voltage  $v_1$ , which is written as  $H_{2_1}$ . This kernel is the opposite of  $H_{2_2}$ . This can be seen in Figure 5.17: the voltage over  $R_1$  in this circuit, which corresponds to  $H_{2_1}$  is the opposite of the voltage over  $R_2$ , which in this circuit corresponds to  $H_{2_2}$ . Hence

$$H_{2_1} = -\frac{1}{(g_1 + g_2)^3} \left( K_{2g_1} g_2^2 - K_{2g_2} g_1^2 \right)$$
 (5.162)

The nonlinear current sources of order three depend on both the first- and second-order kernels of the controlling voltages. Combining the general expression from Table 5.2 for the nonlinear current source of order three corresponding to a nonlinear conductance, with expressions (5.146) and (5.162) yields the third-order nonlinear current source  $i_{NL3g_1}$  that corresponds to  $R_1$ :

$$i_{NL3g_1} = \frac{1}{(g_1 + g_2)^4} \left[ K_{3g_1} g_2^3 (g_1 + g_2) + 2K_{2g_1} K_{2g_2} g_1^2 g_2 - 2K_{2g_1}^2 g_2^3 \right]$$
 (5.163)

Using equations (5.145) and (5.151) we find for the third-order nonlinear current source  $i_{NL3g_2}$  that corresponds to  $R_2$ :

$$i_{NL3_{g_2}} = \frac{1}{(g_1 + g_2)^4} \left[ K_{3g_2} g_1^3 (g_1 + g_2) + 2K_{2g_1} K_{2g_2} g_1 g_2^2 - 2K_{2g_2}^2 g_1^3 \right]$$
 (5.164)

These current sources are now applied in the linearized network as shown in Figure 5.19. From

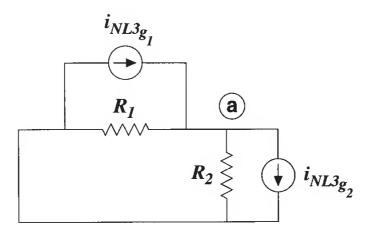


Figure 5.19: Circuit for the computation of the third-order kernels.

this circuit the third-order kernel is computed again by applying Kirchoff's current law at node a. This yields

$$(g_1 + g_2) v_a - i_{NL3g_1} + i_{NL3g_2} = 0 (5.165)$$

Now the unknown voltage  $v_a$  in this equation is the third-order kernel transform of the voltage  $v_2$ , denoted by  $H_{3_2}$ . We find

$$H_{3_2} = \frac{i_{NL3g_1} - i_{NL3g_2}}{g_1 + g_2} \tag{5.166}$$

and using equations (5.163) and (5.164) the third-order kernel of the voltage at node a becomes

$$H_{3_2} = H_{3_a} = \frac{1}{(g_1 + g_2)^5} \left[ K_{3g_1} g_2^3 (g_1 + g_2) - K_{3g_2} g_1^3 (g_1 + g_2) + 2K_{2g_1} K_{2g_2} g_1 g_2 (g_1 - g_2) - 2K_{2g_1}^2 g_2^3 + 2K_{2g_2}^2 g_1^3 \right]$$

$$(5.167)$$

and the third harmonic distortion is found as

$$HD_3 = \frac{V_{in}^2}{4} \cdot \frac{H_{3a}}{H_{1a}} \tag{5.168}$$

It is seen that the third-order kernel becomes zero if the two nonlinearities are identical.

Equation (5.167) is now evaluated for the example of the series connection of the base resistance and  $r_{\pi}$  of a bipolar transistor. The extrinsic part of the base resistance is neglected. For the first- and second-order nonlinearity coefficients the same numerical value are used as before The third-order nonlinearity coefficient that corresponds to  $r_{\pi}$  is found from equation (3.24):

$$K_{3g_{\pi}} = \frac{g_{\pi}}{6V_t^2} = 0.1A/V^3 \tag{5.169}$$

For the third-order nonlinearity coefficient  $K_{3g_{Bi}}$  that describes the third-order behavior of the (intrinsic) base resistance, a value of  $40A/V^3$  is taken. With these value  $H_{3a}$  is equal to  $-56.5V^{-2}$  and  $HD_3$  for an input amplitude of 1V is 15.3. If the base resistance is completely linear, which means that  $K_{2g_{Bi}}$  and  $K_{3g_{Bi}}$  are zero, then  $H_{3a}$  is equal to  $-12.2V^{-2}$ . Hence it is seen that the base resistance again slightly compensates the nonlinearity of  $r_{\pi}$ .

#### 5.5.2 Nonlinear capacitive current divider

Figure 5.20 shows a nonlinear capacitive current divider. The input current  $i_{in}$  is split into the

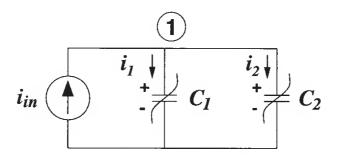


Figure 5.20: A nonlinear capacitive current divider.

currents through the nonlinear capacitors. We will compute the first three harmonics of the current through each capacitor. The harmonics will be computed directly with the method of Section 5.3.

The input current is sinusoidal:

$$i_{in}(t) = I_{in}sin(\omega_1 t) \tag{5.170}$$

The nonlinearity of the capacitors is described as in Section 3.2.3. The current  $i_1$  through capacitor  $C_1$  is given by (see also equation (3.51))

$$i_1(t) = \frac{d}{dt} \left( C_1 \cdot v_1(t) + K_{2C_1} \cdot v_1^2(t) + K_{3C_1} \cdot v_1^3(t) \dots \right)$$
 (5.171)

The current through  $C_2$  is  $i_2$ :

$$i_2(t) = \frac{d}{dt} \left( C_2 \cdot v_1(t) + K_{2C_2} \cdot v_1^2(t) + K_{3C_2} \cdot v_1^3(t) \dots \right)$$
 (5.172)

**First-order responses** First, the linear components of the two currents are determined. To this purpose, the circuit is linearized. The linearized circuit is shown in Figure 5.21. In this circuit it is seen that

$$I_{in} = j\omega_1(C_1 + C_2)V_{1,1,0} (5.173)$$

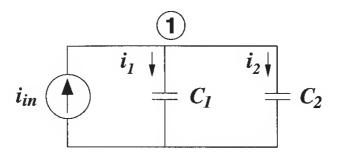


Figure 5.21: Linearized equivalent of the capacitive current divider.

Here  $V_{1,1,0}$  indicates the complex amplitude of the first-order response of the voltage at node 1. The first subscript corresponds to the node number, the second subscript to the order of the harmonic of  $\omega_1$ . The third subscript is zero since we only apply one single frequency.

From equation (5.173)  $V_{1,1,0}$  is found to be

$$V_{1,1,0} = \frac{I_{in}}{j\omega_1(C_1 + C_2)} \tag{5.174}$$

The complex amplitude of the first-order component of the current through capacitor  $C_1$  is given by

$$I_{1,1,0} = V_{1,1,0} \, j\omega_1 C_1 \tag{5.175}$$

or, using equation (5.174)

$$I_{1,1,0} = \frac{C_1}{C_1 + C_2} I_{in}$$
 (5.176)

Similarly, we find for the complex amplitude of the first-order component  $I_{2,1,0}$  of the current through  $C_2$ :

$$I_{2,1,0} = \frac{C_2}{C_1 + C_2} I_{in} (5.177)$$

Second-order responses Next, the second harmonics of the currents  $i_1$  and  $i_2$  are determined. This is performed with the circuit of Figure 5.22. This is the linearized circuit of Figure 5.21 from which the external excitation has been removed. Instead the nonlinear current sources of order two are applied. Since we are computing the second harmonics directly now, the value of the nonlinear current sources must be obtained from Table 5.5. In this way we find:

$$i_{NL2}{}_{C_1} = j\omega_1 K_2{}_{C_1} V_{1,1,0}^2 = j\omega_1 K_2{}_{C_1} \left(\frac{I_{in}}{j\omega_1 (C_1 + C_2)}\right)^2$$
 (5.178)

$$i_{NL2}{}_{C_2} = j\omega_1 K_2{}_{C_2} V_{1,1,0}^2 = j\omega_1 K_2{}_{C_2} \left(\frac{I_{in}}{j\omega_1 (C_1 + C_2)}\right)^2$$
 (5.179)

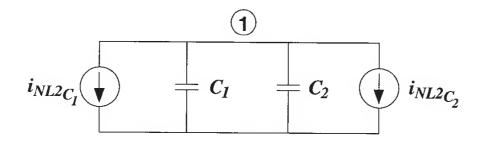


Figure 5.22: Circuit used to compute the second-order responses.

Applying Kirchoff's current law at node 1 in Figure 5.22 yields

$$V_{1,2,0} \cdot 2j\omega_1(C_1 + C_2) = -i_{NL2} - i_{NL2} - i_{NL2}$$
(5.180)

Instead of the node voltage itself, the second-order harmonic of this node voltage has been taken as a node voltage. Also it is seen that the frequency is  $2j\omega_1$ . This is in correspondence with the method for the direct computation of nonlinear responses.

From equation (5.180)  $V_{1,2,0}$  is found

$$V_{1,2,0} = -\frac{i_{NL^2C_1} + i_{NL^2C_2}}{2j\omega_1(C_1 + C_2)}$$
(5.181)

Using equations (5.178) and (5.179) this becomes

$$V_{1,2,0} = -\frac{K_{2C_1} + K_{2C_2}}{2(j\omega_1)^2 (C_1 + C_2)^3} I_{in}^2$$
(5.182)

From Section 5.2.6 we know that the second harmonic of the current through capacitor  $C_1$  is the sum of the current delivered by the nonlinear current of order two and the current through the linear capacitor  $C_1$  in Figure 5.22. Hence we obtain

$$I_{1,2,0} = i_{NL2_{C_1}} + 2j\omega_1 C_1 V_{1,2,0}$$
(5.183)

Using equations (5.178) and (5.182) we find after some algebra

$$I_{1,2,0} = \frac{I_{in}^2}{j\omega_1 \left(C_1 + C_2\right)^3} \left(C_2 K_{2C_1} - C_1 K_{2C_2}\right)$$
 (5.184)

The second harmonic distortion is found by taking the ratio of  $I_{1,2,0}$  and  $I_{1,1,0}$ . This yields

$$HD_{2} = \left| \frac{I_{in}}{j\omega_{1} \left( C_{1} + C_{2} \right)^{2}} \cdot \left( \frac{C_{2}}{C_{1}} K_{2C_{1}} - K_{2C_{2}} \right) \right|$$
 (5.185)

It is seen that  $HD_2$  decreases when the frequency increases. Further, we see that the second harmonic of the current through  $C_1$  is zero if  $C_1$  and  $C_2$  are tracking nonlinearities, since then  $C_2 = aC_1$  and  $K_{2C_2} = aK_{2C_1}$ .

In a similar way the second harmonic of the current through  $C_2$  is found. It is given by

$$I_{2,2,0} = \frac{I_{in}^2}{j\omega_1 \left(C_1 + C_2\right)^3} \left(C_1 K_{2C_2} - C_2 K_{2C_1}\right)$$
 (5.186)

which is seen to be the opposite of  $I_{1,2,0}$ . The second harmonic distortion of the current through  $C_2$  is the same as  $HD_2$  of the current through  $C_1$ .

**Third-order responses** For the computation of the third harmonics the nonlinear current source of order three are determined first. From Table 5.7 we find that the nonlinear current source of order three  $i_{NL_3}$ <sub>C<sub>1</sub></sub> that corresponds to  $C_1$  is given by

$$i_{NL3_{C_1}} = 3j\omega_1 \left[ K_{2_{C_1}} V_{1,1,0} V_{1,2,0} + \frac{1}{4} K_{3_{C_1}} V_{1,1,0}^3 \right]$$
 (5.187)

Using equations (5.174) and (5.182) this becomes

$$i_{NL^{3}C_{1}} = \frac{3}{2} \frac{I_{in}^{3}}{\left(C_{1} + C_{2}\right)^{3} \left(j\omega_{1}\right)^{2}} \left[ \frac{-K_{2C_{1}} \left(K_{2C_{1}} + K_{2C_{2}}\right)}{C_{1} + C_{2}} + \frac{K_{3C_{1}}}{2} \right]$$
(5.188)

Similarly, the nonlinear current of order three corresponding to  $\check{C}_2$  is found to be

$$i_{NL^{3}C_{2}} = \frac{3}{2} \frac{I_{in}^{3}}{\left(C_{1} + C_{2}\right)^{3} \left(j\omega_{1}\right)^{2}} \left[ \frac{-K_{2C_{2}}\left(K_{2C_{1}} + K_{2C_{2}}\right)}{C_{1} + C_{2}} + \frac{K_{3C_{2}}}{2} \right]$$

$$(5.189)$$

The complex amplitude of the third harmonic of the voltage at node 1 is found in the same way as in equation (5.181):

$$V_{1,3,0} = -\frac{i_{NL^3C_1} + i_{NL^3C_2}}{3j\omega_1(C_1 + C_2)}$$
(5.190)

Using equations (5.188) and (5.189) this becomes

$$V_{1,3,0} = -\frac{I_{in}^3}{2\left(C_1 + C_2\right)^4 (j\omega_1)^3} \left[ \frac{K_{3C_1}}{2} + \frac{K_{3C_2}}{2} - \frac{\left(K_{2C_1} + K_{2C_2}\right)^2}{C_1 + C_2} \right]$$
(5.191)

The third harmonic of the current through  $C_1$  is the sum of  $i_{NL^3C_1}$  and the current through the capacitor  $C_1$  in the linearized network that is excited with the nonlinear current sources of order three:

$$I_{1,3,0} = i_{NL^3C_1} + 3j\omega_1 C_1 V_{1,3,0} (5.192)$$

Using equations (5.188) and (5.191) we obtain

$$I_{1,3,0} = \frac{3}{4} \frac{I_{in}^3}{(j\omega_1)^2 (C_1 + C_2)^5} \left[ K_{3C_1} C_2 (C_1 + C_2) - K_{3C_2} C_1 (C_1 + C_2) - 2K_{2C_1}^2 C_2 + 2K_{2C_2}^2 C_1 + 2K_{2C_1} K_{2C_2} (C_1 - C_2) \right]$$
(5.193)

The third harmonic distortion is found by taking the ratio of  $I_{1,3,0}$  with  $I_{1,1,0}$ . One can verify that the third harmonic of the current through  $C_2$  is the opposite of equation (5.193). The third harmonic is zero when  $C_1$  and  $C_2$  are tracking nonlinearities: in that case, the part between the square brackets in the right-hand side of equation (5.193) is indeed zero.

#### 5.6 Numerical verification with other methods

In the previous sections techniques have been discussed to analyze the behavior of a circuit that operates in a weakly nonlinear way. The larger the amplitude of the circuit excitation is, the larger the deviation will be between the weakly nonlinear behavior approximation and the circuit's actual nonlinear behavior. Since it is generally not possible to estimate that deviation analytically, one needs to resort to a numerical simulator to verify the results.

The classical approach for the numerical simulation of nonlinear circuits that are excited by one or more sinusoids, is to perform a time-domain simulation followed by a Fourier transform to see the spectrum of the output signal. With such approach, the voltages and currents are first computed as a function of time by integrating numerically the set of nonlinear differential equations that describe the circuit. Usually, the initial time conditions of the circuit are such that some transient effects occur before the circuit is in a steady state. Hence the time-domain simulation must be performed over a period that is longer than the time required to let the transients die out. The sequence of timepoints that is considered for the Fourier transform must of course begin after the transients have died out. Sometimes, these transients remain in effect for a long time, for example in high-Q filters.

Numerical integration has some drawbacks, as will be discussed in Section 5.6.1. Two major alternatives of this numerical integration are used in circuit simulation: shooting methods, which are discussed briefly in Section 5.6.2, and harmonic balance methods, discussed in Section 5.6.3. An excellent survey of these methods can be found in [Kund 90].

## 5.6.1 Numerical integration

Numerical integration is used for example in SPICE with a so-called transient analysis. Although widely used, the classical numerical integration has several drawbacks. If, for example, the time constants in the circuit are much larger than the period of the excitation, then a lot of integration cycles need to be computed. Apart from the extra CPU time, this situation can also lead to an accumulation of roundoff errors.

Another example for which numerical integration becomes impractical is the simulation of an upconversion or downconversion mixer. Typically, the RF frequency is several orders of

magnitude higher than the baseband frequency or frequencies. Again, the number of timesteps will be huge: the time interval over which the differential equations have to be integrated is determined by the lowest frequency, whereas the size of the timestep is determined by the highest frequency response of the circuit.

In order to overcome these problems, the methods described in the following sections can be used.

#### 5.6.2 Shooting methods

Shooting methods integrate the circuits differential equations over several intervals of one period. On each iteration, the initial conditions are varied, trying to match the signals at the end of the period with those at the beginning. If they match, then those initial conditions are found that do not cause transients, and so the steady state is found. With some modifications, this principle can be extended to compute responses to more than one sinusoidal signal. Such responses are denoted as almost periodic solutions. More details about shooting methods can be found in the original papers [Apr 72, Skel 80] or in text books that present an overview [Vlach 83, Kund 90].

The advantage of shooting methods over classical numerical integration is that the total number of timepoints, which is the number of timepoints in one period, times the number of required periods, is smaller. For example, the shooting method that has been implemented in [Dam 93] yields a reduction of the CPU time with a factor 6, compared to numerical integration.

#### 5.6.3 Harmonic balance methods

When the steady-state response of a nonlinear circuit is computed using a numerical integration rule, then the steady-state solution is constructed as a collection of time samples with an implied interpolating function. Typically, the interpolating function is a low-order polynomial. Assume now that the steady-state solution consists of a sum of sinusoids. For a good approximation of such solution using numerical integration, many time points are required, since polynomials fit sinusoids poorly.

Harmonic balance methods, on the other hand, use a linear combination of sinusoids to build the solution. This is especially advantageous when the steady-state response contains sinusoids at widely separated frequencies.

A representation of the steady-state solution as a series of sine and cosine functions has also been used in the method to directly compute harmonics and intermodulation products, described in Section 5.3 and Appendix C. This calculation method is restricted to weakly nonlinear behavior, whereas the harmonic balance methods are able to simulate strongly nonlinear behavior as well. In Appendix C it is also made clear that with a representation of the steady-state solution as a sum of sine and cosine functions, dynamic operations such as differentiation and integration reduce to a multiplication with frequency and a division by the frequency, respectively. In this way, nonlinear differential equations reduce to nonlinear algebraic equations.

With the calculation method explained in Section 5.3 the nonlinearities are described by polynomials of at most degree three, and all responses due to those nonlinearities of order higher than three are neglected. With these assumptions the coefficients of the sine and cosine functions

are computed. Harmonic balance methods, on the other hand, can handle arbitrary nonlinearities, that in general cannot be described accurately by a polynomial of degree three. In this case the coefficients of the sine and cosine functions that represent the steady-state solution, are computed by converting into the time domain and then back again [VdEi 89, Kund 90].

In order to describe strongly nonlinear behavior of a circuit in an accurate way it is clear that many sine and cosine functions will have to be taken into account. Since every node voltage and possibly some branch currents have to be described by a long series of sine and cosine functions, it is clear that harmonic balance methods require many memory resources for the computations. The situation even becomes worse when the nonlinear circuit is excited by more than one frequency, which is the case in mixers. However, in recent years drastic improvements in the efficiency of harmonic balance methods have been reported with the so-called Krylov subspace approach, allowing for example to simulate complete analog front-end chips at the transistor level with the harmonic balance method [Feld 96].

#### 5.6.4 Example: an emitter follower

The validity of the low-distortion conditions that are implicitly assumed in the method of Section 5.3 for the direct calculation of nonlinear responses, is now checked for an emitter follower. This circuit is shown in Figure 8.47. The collector current of transistor  $Q_1$  is 1mA and the value of  $R_E$  is  $1k\Omega$ . For this circuit the first three harmonics at the emitter are computed both with the calculation method of Section 5.3 and with a harmonic balance method [Kund 90].

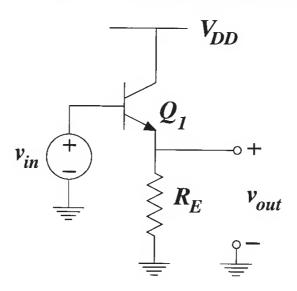


Figure 5.23: An emitter follower.

Figure 5.24 shows the three harmonics at the output of the emitter follower of Figure 5.23 as a function of the amplitude of the input voltage source, which is a sinusoid at low frequencies. With the calculation method of Section 5.3, the fundamental response increases proportionally with the input amplitude, the second harmonic with the square and the third harmonic with the third power of the input amplitude. This is an approximation that clearly is not valid anymore

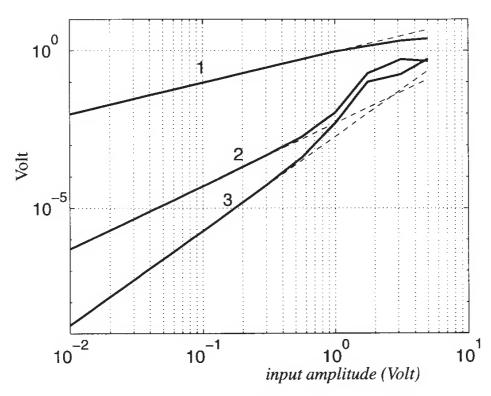


Figure 5.24: The first three harmonics at the emitter of  $Q_1$  as a function of the input amplitude, computed with a harmonic balance method (solid line) and with the calculation method of Section 5.3 (dashed line).

at large input amplitudes. Nevertheless the approximation is very good until amplitudes as high as 400mV which is 40% of the product  $I_CR_E$ . Hence, the method of Section 5.3 is sufficiently accurate up to 40% of the full signal swing. Moreover, this method is very efficient since it only requires the solution of sets of linear equations. On the other hand, the numerical techniques described in this section require iteration and thus more computer resources in general.

## 5.7 Summary

In this chapter we explained how closed-form expressions can be obtained for the nonlinear behavior of analog integrated circuits. In general, it is not possible to obtain closed-form expressions for a circuit's nonlinear behavior. Many techniques that are used to compute the nonlinear behavior numerically, such as the harmonic balance method [VdEi 89, Kund 90], would give problems when they would be used to obtain closed-form expressions. The reason is that such techniques use iteration. If iterations would be performed symbolically, then — in the best case — this would yield large, uninterpretable expressions.

Two methods have been explained and illustrated for the computation of nonlinear responses (harmonics and intermodulation products). The methods are very closely related. The first method computes Volterra kernel transforms. This is performed by repeatedly solving a lin-

ear circuit. From the knowledge of the kernel transforms, the wanted nonlinear responses can be computed using the relationships between Volterra kernel transforms and nonlinear responses.

With the other method, the nonlinear responses are directly computed. The use of Volterra kernels is circumvented here, which is a significant simplification for multiple-input circuits. Just as with the Volterra series approach, the method computes harmonics or intermodulation products of voltages and currents in a weakly nonlinear circuit by repeatedly solving a linear circuit.

Since with both calculation methods a linear circuit is solved, it is possible to obtain symbolic expressions which are functions of the linear small-signal parameters and of the nonlinearity coefficients that explicitly express the nonlinear nature of the devices. These nonlinearity coefficients have been discussed in Chapter 3.

With the methods explained in this chapter, a harmonic or intermodulation product is computed as a sum of several contributions, one for each nonlinearity coefficient. Since in practical circuits many contributions occur, it is impractical to generate a fully symbolic expression for the harmonic or intermodulation product. A procedure has been explained that discards the insignificant contributions with a user-defined error before the symbolic computations. After the application of this procedure usually many nonlinearities can be discarded. Then a symbolic expression can be generated with the few remaining nonlinearities. The resulting symbolic expression is a hierarchical one, that can be pruned. This results into an interpretable expression that allows to obtain insight in a circuit's nonlinear behavior. This approach will be illustrated in Chapter 8 with transistor networks. The nonlinearity coefficients used in the calculations there, will be discussed in the next two chapters for bipolar and MOS transistors.

# Chapter 6

# Silicon bipolar transistor models for distortion analysis

#### 6.1 Introduction

In this chapter the most important nonlinearities in a bipolar transistor are discussed. The transistor is assumed to work in the forward active region. The model that will be followed in this chapter is the Gummel-Poon model [Getr 76]. This model is used now already for several decades. More recent models such as the VBIC95 model [Mc And 95] are still based upon this model. The region of quasi-saturation [Anto 88] is not considered here. This region is especially of interest for power transistors. Distortion caused by quasi-saturation in power applications is discussed in [DeVr 96].

The equivalent nonlinear circuit for a bipolar transistor that will be used throughout this book is shown in Figure 6.1. It consists of lumped elements only, although in reality the circuit

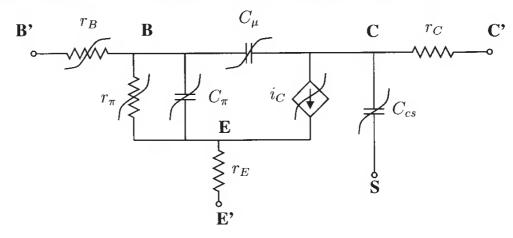


Figure 6.1: Nonlinear model of a bipolar transistor.

elements have a distributed nature [Taft 91].

The base-emitter diode (between **B** and **E** in Figure 6.1), which is forward biased, is represented by a nonlinear conductance  $r_{\pi}$  in parallel with a junction capacitance and a diffusion

capacitance. The sum of these two capacitances is denoted as  $C_{\pi}$ . Since the diffusion capacitance at a sufficiently high forward bias is much larger than the junction capacitance, the nonlinearity coefficients of the latter are neglected. The base-collector junction (between **B** and **C** in Figure 6.1) is reversely biased. It is represented by a junction capacitance only, denoted as  $C_{\mu}$ , which is modeled in the same fashion as described in Section 3.4.

The collector current, which flows between collector and emitter, is a function of two voltages, the base-emitter voltage and the base-collector voltage. The dependence on the latter voltage in the forward active operation is only due to the Early effect. In the linearized equivalent circuit of a transistor, the Early effect is modeled by taking into account a resistance between collector and emitter. This looks as if the collector current would depend on the base-emitter voltage and the collector-emitter voltage. This convention is followed here as well, whereas the Gummel-Poon model assumes a dependence of the collector current on  $v_{BE}$  and  $v_{BC}$ . However, by substituting  $v_{BC}$  with  $(v_{BE}-v_{CE})$  in the Gummel-Poon equations, the collector current can be expressed as a function of  $v_{BE}$  and  $v_{CE}$ .

The Early effect is usually modeled by assuming a linear dependence of the collector current on the base-collector or collector-emitter voltage. This is only an approximation. More accurate models are given in this chapter or in [Joard 95, Mc And 96]. Measurements on the nonlinearity of the Early effect are presented in Chapter 9.

Three ohmic resistors are considered in the equivalent scheme. They connect the "intrinsic" transistor to the outside world via the terminals  ${\bf B'}$ ,  ${\bf E'}$  and  ${\bf C'}$ . As will be discussed below, the base resistance  $r_B$  is a nonlinear resistance that depends on the base current through the effects of current crowding, base conductivity modulation and base pushout [Yuan 88, Sata 90, Fuse 95]. The ohmic resistors  $r_C$  and  $r_E$  are considered as linear elements. This is not exact, as explained in [Taft 91, Fuse 95]: they decrease slightly with increasing bias current. However, their variation as a function of signal swing is not as drastic as for the base resistance, if at least the transistor is assumed to remain in the forward active region. It is believed that the importance of the resistive parasitics will increase for future technologies. Moreover, as the device current increases with scaling, an accurate modeling of the variation of these resistive parasitics with bias currents will become more and more important.

The last circuit element to be considered in the equivalent circuit of Figure 6.1 is the collector-substrate junction capacitance  $C_{cs}$ . This capacitance is modeled as described in Section 3.4.

Finally, it should be noted that frequency effects due to the distributed nature of stored charge in the base and emitter, are not taken into account with the simple model of Figure 6.1. The influence of non-quasi-static charge distributions in base and emitter is modeled for example in [Hamel 96]. These influences can be modeled by a circuit with lumped elements which is more complicated than the linearized equivalent of the equivalent circuit of Figure 6.1. A first attempt for a better high-frequency modeling, which is also implemented in SPICE, is a split of  $C_{\mu}$  into two parts [Anto 88, Hspi 96], one part between B' and C, and the other part between B and C.

#### 6.2 The collector current

According to the Gummel-Poon model [Getr 76, Anto 88], the collector current  $i_C$  can be considered as a function of two voltages,  $v_{BE}$  and  $v_{CE}$ . The dependence on  $v_{CE}$  is due to the Early effect. This effect will be explained in Section 6.2.2. First, this dependence will be considered as linear.

#### 6.2.1 Collector current with a linear Early effect

When high injection is not taken into account and the Early effect is linear, then the collector current  $i_C$  is given by [Getr 76, Gray 93, Lak 94]

$$i_C = I_S \exp\left(\frac{v_{BE}}{n_F V_t}\right) \left(1 + \frac{v_{CB}}{V_{AF}}\right) \tag{6.1}$$

$$=I_S \exp\left(\frac{v_{BE}}{n_F V_t}\right) \left(1 + \frac{v_{CE} - v_{BE}}{V_{AF}}\right) \tag{6.2}$$

in which  $I_S$ ,  $n_F$ ,  $V_{AF}$  and  $V_t$  are the transistor saturation current, the forward emission coefficient, the forward Early voltage and the thermal voltage, respectively. The forward emission coefficient is equal to one, in theory, but it is sometimes fitted to a slightly different value.

The total value of the collector current is the sum of the quiescent value  $I_C$  and the AC value  $i_c$ . A Taylor series expansion around the quiescent point  $I_C$ ,  $V_{BE}$ ,  $V_{CE}$  leads to the expression for the AC current given in equation (3.61), which is repeated here for convenience:

$$i_{c} = g_{m} \cdot v_{be} + K_{2g_{m}} \cdot v_{be}^{2} + K_{3g_{m}} \cdot v_{be}^{3} + \dots + g_{o} \cdot v_{ce} + K_{2g_{o}} \cdot v_{ce}^{2} + K_{3g_{o}} \cdot v_{ce}^{3} + \dots + K_{2g_{m} \& g_{o}} \cdot v_{be} \cdot v_{ce} + K_{3g_{m} \& 2g_{o}} \cdot v_{be} \cdot v_{ce}^{2} + \dots$$

$$(6.3)$$

The meaning of the coefficients in this equation, together with their value that has been derived from equation (6.2), is given in Table 6.1. In Table 6.1 it is seen that coefficients that comprise derivatives with respect to  $v_{CE}$  of order higher than one, are zero. This is due to the modeling of the Early effect in a linear way. More realistic values will be obtained in the next section.

From the values of Table 6.1 we find the second- and third-order normalized coefficients that describe the nonlinear dependence of the collector current on  $v_{BE}$ :

$$K_{2g_m}' = K_{2g_m}/g_m = \frac{1}{2n_F V_t} \tag{6.4}$$

$$K_{3g_m}' = K_{3g_m}/g_m = \frac{1}{6n_F^2 V_t^2}$$
 (6.5)

It is seen that these normalized nonlinearity coefficients are independent of bias conditions.

The normalized nonlinearity coefficients for the dependence of the collector current on  $v_{BE}$  are very large. For example, in Chapter 3, Section 3.3, it has been mentioned that the normalized second-order nonlinearity coefficient of a diffused resistor is about 0.5%/V. This is very low compared to the value of  $K'_{2g_m}$ , which is about 2000%/V at room temperature!

$g_m$	$rac{\partial i_C}{\partial v_{BE}}$	$rac{I_C}{n_F V_t}$
$K_{2g_m}$	$\frac{1}{2} \frac{\partial^2 i_C}{\partial v_{BE}^2}$	$rac{g_m}{2n_FV_t}$
$K_{3g_m}$	$\frac{1}{6} \frac{\partial^3 i_C}{\partial v_{BE}^3}$	$\frac{g_m}{6n_F^2V_t^2}$
$g_o$	$rac{\partial i_C}{\partial v_{CE}}$	$rac{I_C}{V_{AF}}$
$K_{2g_o}$	$\frac{1}{2} \frac{\partial^2 i_C}{\partial v_{CE}^2}$	0
$K_{3g_o}$	$\frac{1}{6} \frac{\partial^3 i_C}{\partial v_{CE}^3}$	0
$K_{2_{g_m}\&g_o}$	$rac{\partial^2 i_C}{\partial v_{BE} \partial v_{CE}}$	$rac{g_m}{V_{AF}}$
$K_{3_{2g_m\&g_o}}$	$\frac{1}{2} \frac{\partial^3 i_C}{\partial v_{BE}^2 \partial v_{CE}}$	$\frac{g_m}{2n_F V_t V_{AF}}$
$K_{3_{g_m\&2g_o}}$	$\frac{1}{2} \frac{\partial^3 i_C}{\partial v_{BE} \partial v_{CE}^2}$	0

Table 6.1: Nonlinearity coefficients of the collector current of a bipolar transistor obtained with the model of equation (6.2).

At high collector currents, the injection of minority carriers into the base region is significant with respect to the majority carrier concentration in equilibrium. Since charge neutrality is maintained in the base, the total majority carrier concentration is increased by the same amount as the total minority carrier concentration. In [Getr 76, Lak 94] it is shown that at high injection levels the collector current asymptotes to

$$i_{C\, ({\rm high \, level})} \propto \exp\left(\frac{v_{BE}}{2V_t}\right)$$
 (6.6)

In the Gummel-Poon model a unified equation describes the collector current, both for low and

high injection levels [Getr 76, Anto 88, Hspi 96]:

$$i_C = \frac{2I_S \exp\left(\frac{v_{BE}}{n_F V_t}\right) \left(1 + \frac{v_{CE} - v_{BE}}{V_{AF}}\right)}{1 + \sqrt{1 + \frac{4I_S \exp\left(\frac{v_{BE}}{n_F V_t}\right)}{I_{KF}}}}$$

$$(6.7)$$

in which  $I_{KF}$  is the forward knee-current [Getr 76, Anto 88, Lak 94]. This current is the value of the collector current at which high injection begins: for currents that are much smaller than  $I_{KF}$ , the logarithm of the collector current increases as a function of  $v_{BE}$  with a slope  $1/(n_F V_t)$ . For currents much higher than  $I_{KF}$  the slope is  $1/(2n_F V_t)$ . The value of  $I_{KF}$  can now be found as the current where the asymptotes with the two slopes intersect.

Next, the nonlinearity coefficients are derived with the model of equation (6.7). Under high-injection conditions the collector current is proportional to  $\exp(v_{BE}/(2n_FV_t))$ , such that

$$g_m ext{ (high injection)} = \frac{i_C}{2n_F V_t} ext{ (6.8)}$$

$$K_{2g_m}$$
 (high injection) =  $\frac{g_m \text{ (high injection)}}{4n_F V_t} = \frac{i_C}{8n_F^2 V_t^2}$  (6.9)

$$K_{3g_m}$$
 (high injection) =  $\frac{g_m \text{ (high injection)}}{24n_F^2 V_t^2} = \frac{i_C}{48n_F^3 V_t^3}$  (6.10)

and for the normalized nonlinearity coefficients

$$K'_{2g_m}$$
 (high injection) =  $\frac{1}{4n_FV_t}$  (6.11)

$$K_{3g_m}'$$
 (high injection) =  $\frac{1}{24n_F^2V_t^2}$  (6.12)

Compared to the low-injection situation (see Table 6.1) the normalized second-order coefficient is a factor 2 lower, while the third-order normalized nonlinearity coefficient is a factor 4 lower.

Figures 6.2 and 6.3 show the normalized nonlinearity coefficients of order two and three as a function of the collector current. The parameters for these plots are:  $I_S = 6.5 \times 10^{-18} A$ ,  $n_F = 1.0$  and  $I_{KF} = 8.0 \times 10^{-3}$ . It is seen that the normalized nonlinearity coefficients vary between the low-injection value (equations (6.4) and (6.5)) and the high-injection value (equations (6.11) and (6.12)). The value of the normalized coefficients is seen to decrease already at currents that are only a small fraction of  $I_{KF}$ .

#### 6.2.2 Nonlinearity of the Early effect

The Early effect is due to the shortening of the neutral base-width as the collector-base voltage increases. By this increase, the base-collector junction gets more inversely biased, such that the

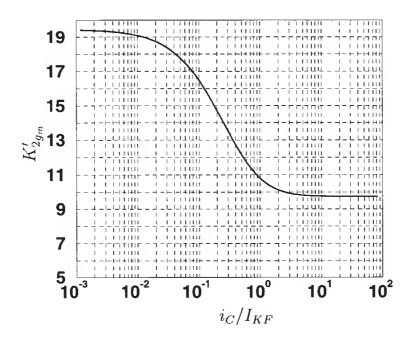


Figure 6.2: Second-order normalized nonlinearity coefficient  $K'_{2g_m}$  as a function of the collector current which is normalized to the forward knee current  $I_{KF}$ .

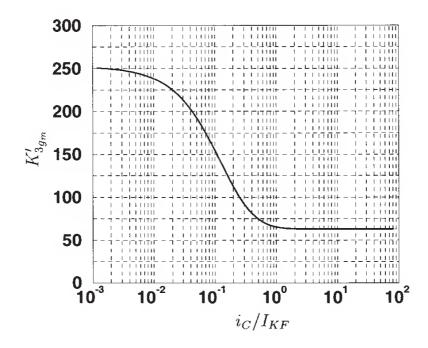


Figure 6.3: Third-order normalized nonlinearity coefficient  $K'_{3g_m}$  as a function of the collector current which is normalized to the forward knee current  $I_{KF}$ .

space-charge extends more in the neutral base. As a result, the length of this neutral region is reduced and the majority charge in the neutral base region, denoted by  $Q_B$ , decreases. This is shown schematically in Figure 6.4. The extension of the space-charge layer into the neutral

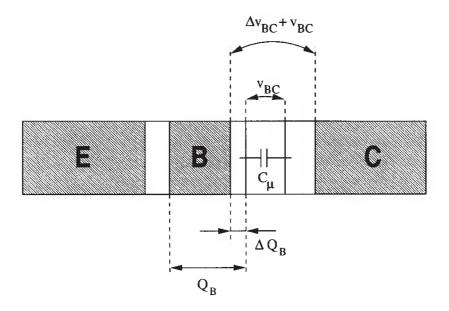


Figure 6.4: Decrease of  $Q_B$  with an increase of  $v_{CE}$  or  $v_{CB}$ . The hatched regions represent the neutral regions. The white regions are the space-charge layers.

collector region is higher than the extension into the neutral base. This is due to the higher doping of the collector.

Since the saturation current  $I_S$  is inversely proportional to  $Q_B$  [Getr 76, Lak 94], the collector current increases as  $v_{CE}$  increases. This increase can be modeled with an output conductance  $g_o$ :

$$g_o = \frac{di_C}{dv_{CE}} \tag{6.13}$$

This conductance is calculated as follows. The variation of the collector-emitter voltage  $\Delta v_{CE}$  can be rewritten as  $\Delta v_{CE} = \Delta v_{CB} + \Delta v_{BE}$ . This is equal to  $\Delta v_{CB}$ , since the base-emitter voltage is kept constant. Equation (6.13) can then be rewritten as

$$g_o = \frac{di_C}{dv_{CB}} = \frac{di_C}{dQ_B} \cdot \frac{dQ_B}{dv_{CB}} \tag{6.14}$$

The dependence of  $i_C$  on  $Q_B$  is through the saturation current  $I_S$ , which is inversely proportional to  $Q_B$ . The change of  $Q_B$  with respect to  $v_{CB}$  is nothing else but  $-C_\mu$  since a slice of neutral base is replaced by a slice of space-charge layer. Hence,

$$g_o = -\frac{i_C}{Q_B} \left( -C_{\mu} \right) = \frac{i_C C_{\mu}}{Q_B} \tag{6.15}$$

When this equation is evaluated for  $v_{CB} = 0V$ , then we obtain

$$g_o \Big|_{v_{CB}=0} = \frac{i_C}{Q_{B0}} C_\mu \Big|_{v_{CB}=0}$$
 (6.16)

in which  $Q_{B0}$  is the majority base charge at  $v_{CB}=0V$ . The ratio  $C_{\mu}/Q_{B0}$  has the dimensions of a voltage. This voltage is called the *Early voltage*  $V_{AF}$ . It is given by

$$V_{AF} = \frac{C_{\mu}|_{v_{CB}=0}}{Q_{B0}} \tag{6.17}$$

In most circuit simulators it is assumed that for the modelling of the Early effect, the basecollector capacitance  $C_{\mu}$  does not change with bias [Getr 76, Anto 88, Mc And 96], such that  $V_{AF}$  can be considered as a constant. This assumption yields a linear dependence of the collector current on  $v_{CE}$ . In this way it is impossible to predict second- and third-order derivatives of the collector current with respect to  $v_{CE}$ . If, on the other hand, the variation of  $C_{\mu}$  is taken into account to model the Early effect, then the nonlinearity coefficients obtained in this way, correspond to measurements as will be described in Chapter 9. For a computation of the nonlinearity coefficients the reader is referred to Appendix D. The expressions of the nonlinearity coefficients are quite involved. Nevertheless, their value can be predicted qualitatively. From the discussion about the nonlinearity of junction capacitors in Section 3.4 it is clear that a junction capacitor behaves more and more as a linear capacitor when the reverse bias of a junction increases. In other words, a junction capacitance becomes more independent of the reverse bias when the latter increases. As a consequence, the variation of the collector current with  $v_{CE}$  becomes smaller as  $v_{CE}$  increases. This means that the output conductance will decrease when  $v_{CE}$  increases. The output conductance as a function of  $v_{CE}$  is shown in Figure 6.5. This plot is obtained with the following numerical values:  $I_S = 6.5 \times 10^{-18}$ ,  $C_{\mu} = 28 fF$  at  $v_{BC} = 0V$ ,  $I_{KF} = 8 mA$  and  $V_{AF} = 30V$ . The collector current at  $V_{BC} = 0V$  is 0.85mA.

Since the output conductance decreases when  $V_{CE}$  increase, the second-order derivative of the collector current with respect to  $V_{CE}$  is negative, such that  $K_{2g_o} < 0$ . Figure 6.6 depicts the second-order nonlinearity coefficient  $K_{2g_o}$  as a function of  $V_{CE}$ . The third-order coefficient  $K_{3g_o}$  is shown as a function of  $V_{CE}$  in Figure 6.7.

It is interesting to have an idea of the order of magnitude of the normalized nonlinearity coefficients. Figure 6.8 depicts the normalized nonlinearity coefficients  $K'_{2g_o}$  and  $K'_{3g_o}$ . It is seen that the Early effect is in fact "quite linear". The output conductance has a voltage coefficient that varies between 15% at low values of  $V_{CE}$  and 3% at high values.

Next, the cross-derivatives of the collector current with respect to both  $v_{CE}$  and  $v_{BE}$  are considered. Since it is seen that the output conductance is quite linear, it is reasonable to consider the Early voltage as being constant for the computation of nonlinearity coefficients that correspond to cross-derivatives in which only one differentiation with respect to  $v_{CE}$  is performed. Hence, all values of Table 6.1 are good approximations — at least at low injection— for the nonlinearity coefficients that are not zero in that table. The results of this section provide good values for the nonlinearity coefficients that are zero in Table 6.1. The coefficient  $K_{3_{g_m\&2g_o}}$  corresponds to the only cross-derivative in which two differentiations with respect to  $v_{CE}$  are performed. Its value

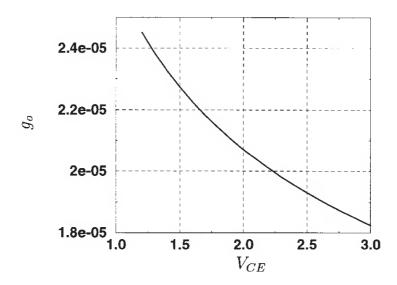


Figure 6.5: The output conductance  $g_o$  as a function of  $V_{CE}$ .

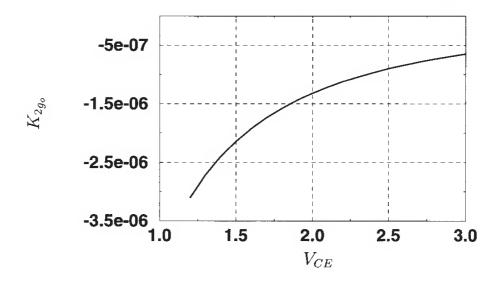


Figure 6.6: Second-order nonlinearity coefficient  $K_{2g_o}$  as a function of  $V_{CE}$ .

under low injection conditions can be obtained by dividing the value of  $K_{2g_o}$  by the thermal voltage  $V_t$ .

#### 6.3 The base current

In the forward active region of operation, the base current is given by [Anto 88]

$$i_B = \frac{I_S}{\beta_F} \exp\left(\frac{q \, v_{BE}}{n_F \, kT}\right) + I_{SE} \exp\left(\frac{q \, v_{BE}}{n_E \, kT}\right) \tag{6.18}$$

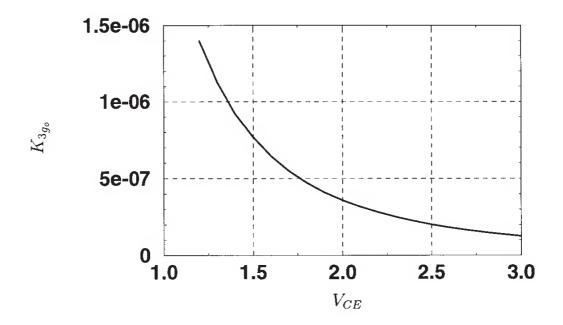


Figure 6.7: Third-order nonlinearity coefficient  $K_{3g_o}$  as a function of  $V_{CE}$ .

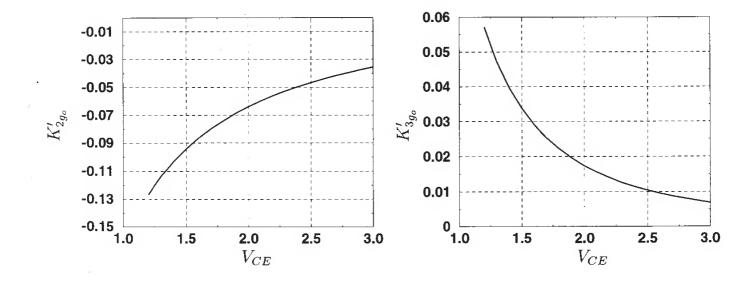


Figure 6.8: Normalized nonlinearity coefficients  $K'_{2g_o}$  and  $K'_{3g_o}$  as a function of  $V_{CE}$ .

The first term is a factor  $\beta_F$  smaller than the collector current if for the latter high injection is neglected. The second term of this expression is caused by recombination and is only important at very low base currents:  $I_{SE}$  is denoted as the base-emitter leakage saturation current, and  $n_E$  is the base-emitter emission coefficient, which is close to 2.

For the determination of the nonlinearity coefficients, equation (6.18) is developed into a power series around the quiescent point, and the AC part is identified with the series

$$i_b = g_{\pi} \cdot v_{be} + K_{2g_{\pi}} \cdot v_{be}^2 + K_{2g_{\pi}} \cdot v_{be}^3 + \dots$$
 (6.19)

in which

$$g_{\pi} = \frac{di_B}{dv_{BE}} \tag{6.20}$$

$$K_{2g_{\pi}} = \frac{1}{2} \cdot \frac{d^2 i_B}{dv_{BE}^2} \tag{6.21}$$

$$K_{3g_{\pi}} = \frac{1}{6} \cdot \frac{d^3 i_B}{dv_{BE}^3} \tag{6.22}$$

At very low base currents, one obtains

$$g_{\pi} = \frac{i_B}{n_E V_t} \tag{6.23}$$

$$K_{2g_{\pi}} = \frac{g_{\pi}}{2n_E V_t} \tag{6.24}$$

$$K_{3g_{\pi}} = \frac{g_{\pi}}{6n_E^2 V_t^2} \tag{6.25}$$

while in the mid-current region we find

$$g_{\pi} = \frac{i_B}{n_F V_t} \tag{6.26}$$

$$K_{2g_{\pi}} = \frac{g_{\pi}}{2n_F V_t} \tag{6.27}$$

$$K_{3g_{\pi}} = \frac{g_{\pi}}{6n_F^2 V_t^2} \tag{6.28}$$

From the above expressions of the nonlinearity coefficients, it is seen that in the mid-current region, and under the assumption that no high injection in the collector region occurs, the nonlinearities of the base current and of the collector current are tracking nonlinearities. Their nonlinearity coefficients only differ by a constant factor  $\beta_F$  and the normalized nonlinearity coefficients are equal. At very low base currents and at high collector currents this is no longer true.

#### 6.4 The base resistance

In modern bipolar processes, the role of the base resistance becomes more and more important: as the base-width becomes smaller, the section through which the carriers have to flow in the base region becomes smaller, such that the base resistance increases. The base resistance is not constant. It decreases when the base current increases. Due to this current dependence, the value of the base resistance changes when the base current changes due to a sinusoidal excitation. This change in base resistance can cause distortion. This distortion can be modeled with the nonlinearity coefficients of the base resistance. These coefficients are discussed in this section.

It is seen that the current density in modern bipolar transistors tends to increase. As a result, the effects that cause the base resistance to decrease, occur at lower currents than in older bipolar processes. This is especially important in high-frequency applications, since bias currents are usually high in these applications.

The base resistance is a component that is difficult to model and difficult to measure. There are several effects that cause the base resistance to decrease as the base current increases: current crowding, base conductivity modulation, base-width modulation and base pushout [Lak 94, Taft 91, Yuan 88, Fuse 92, Fuse 95]. These effects are now shortly discussed.

Current crowding This effect is best explained using Figure 3.9 of Chapter 3. As the base current flows through the active base region, a potential drop in the horizontal direction causes a progressive lateral reduction of the forward bias along the emitter-base junction. In other words, the diodes in the representation of Figure 3.9 close to x=0 are less conducting than the diodes close to x=L. As a result, the current is concentrated close to x=L and the resistive path through the base has become shorter. In the limit, the intrinsic base resistance is bypassed completely and the only base resistance left is the extrinsic base resistance.

**Base-width modulation** In Section 6.2.2 it has been pointed out that the quasi-neutral base-width changes due to a change of the extension of the space-charge layer between base and collector. The same is true for the extension of the space-charge layer between base and emitter [Yuan 88]. The modulation of both of these space-charge layers with the terminal voltages of the transistor affects the base resistance. The reason is that the intrinsic base resistance is inversely proportional to the width of the quasi-neutral base region. For example, assume that the emitter-base junction is forwardly biased and the collector-base voltage remains constant. Then the width of the neutral base region is modulated by the moving edge of the emitter-base space-charge region as  $v_{BE}$  changes. When  $v_{BE}$  increases, the emitter-base space-charge region shrinks and the neutral base region widens. This reduces  $R_{Bi}$ , since the current can flow now through a wider cross-section.

Base conductivity modulation When an *npn* bipolar transistor operates at high currents, the hole concentration (including the excess carrier concentration) in the base will exceed the concentration of the acceptors in order to maintain charge neutrality. As a result, the base sheet resistance below the emitter decreases as the hole injection level increases. Hence, the base resistance decreases.

**Base pushout** At low currents it is assumed that the charge in the space-charge layer between base and collector is only caused by the fixed ion charges, the charge of the mobile carriers being negligible. At high currents, however, this assumption is no longer valid. In an npn transistor, the negative charge of the electrons is no longer negligible compared to the positive charge of the fixed donor ions in the space-charge layer at the collector side. The two charges have an opposite sign, such that together they cause a decrease of the electric field in the collector region near the base. At very high currents, the electrical field becomes very small near the junction. This effect,

in conjunction with high currents, has been studied for example in [Whitt 69]. There it is found that at very high currents, the effective neutral base region extends into the collector region. This corresponds to an increase of the cross-sectional area through which the base current can flow, thereby lowering the base resistance.

## 6.4.1 Modeling of the current dependence

In most versions of SPICE [Anto 88, Hspi 96] the model of the base resistance only takes into account current crowding. The value of the resistance is derived by starting from Figure 3.9. This figure shows that the base resistance is distributed over the base-emitter diode. From this figure it is seen that  $v(L) = v_{BE}^{-1}$ . In order to obtain a more practical model, this distributed model is represented by an equivalent model which consists of a series connection of two lumped elements, a diode and a (nonlinear) resistor, which corresponds to the intrinsic base resistance  $R_{Bi}$ . The extrinsic base resistor  $R_{Bex}$  should be placed in series with  $R_{Bi}$ .

With the model used in SPICE, the effective intrinsic base resistance  $R_{Bi}$  is chosen to dissipate the same power as the distributed intrinsic base resistance for the same base current:

$$\int_{0}^{L} i_{B}^{2}(x) \frac{R_{BiT}(x)}{L} dx = R_{Bi} \cdot i_{B}(L)$$
 (6.29)

in which  $R_{BiT}$  is the total resistance of the resistor chain in the left part of Figure 3.9. This leads to the following expression for the intrinsic base resistance [Anto 88]:

$$R_{Bi} = R_{BiT} \frac{\tan Z - Z}{Z \tan^2 Z} \tag{6.30}$$

in which  $Z \tan Z$  is a function of the base current, the base resistivity, the thermal voltage and the width of the base region. In SPICE [Anto 88, Hspi 96] Z is approximated by

$$Z \approx \frac{-1 + \sqrt{1 + 144i_B/(\pi^2 I_{RB})}}{(24/\pi^2)\sqrt{i_B/I_{RB}}}$$
(6.31)

in which the model parameter  $I_{RB}$  is the current at which the intrinsic base resistance falls down to half the low-current value. At very low currents, Z approaches 0 and, according to equation (6.30),  $R_{Bi}$  approaches  $R_{BiT}/3$ . At high currents, Z approaches  $\pi/2$  and  $R_{Bi}$  goes to zero. For a single-sided base contact, one can show [Taft 91, Chiu 92, Lak 94] that at very low base currents indeed  $R_{Bi} = R_{BiT}/3$ , whereas for double base contacts  $R_{Bi} = R_{BiT}/12$ .

Since the above model has been derived using power dissipation considerations, this model is often referred to as the *power model* for a base resistance.

In [Taft 91, Chiu 92] it is shown that the above reasoning leads to incorrect values for the base resistance. This is not due to the nonlinearity of the circuit elements but rather to the nonlinearity of power as a function of current. Also, equation (6.30) is used in SPICE not only for the DC

the emitter resistance is neglected in this figure.

base resistance but also for the AC base resistance. In [Taft 91, Chiu 92, Fuse 95], however, it is shown that the AC value of the intrinsic base resistance differs from the DC value.

A better model for current crowding is presented by Taft and Plummer [Taft 91], which has later been improved by Chiu [Chiu 92]. The latter model takes into account base conductivity modulation by the excess carriers. The two models have been verified in [Asti 96], where it is shown that the model of Chiu is more accurate at higher currents, whereas the accuracy with both models is good at moderate currents. In Section 6.4.3 both the model of Chiu and the power model will be used for the determination of the nonlinearity coefficients.

Other models The different effects of current crowding, base width modulation, base conductivity modulation and base pushout can occur simultaneously. In fact, base width modulation, conductivity modulation and pushout ameliorate the effect of current crowding. A model for the base resistance that takes into account all these effects is very complicated. Such model is described for example in [Yuan 88]. This model is no longer an analytic one, it must be obtained by numerical integration techniques. In this model the base resistance is given by the following equation

$$R_{Bi} = R_{Bi0} \cdot f_{BWM} \cdot f_{CM} \cdot f_{CROWDING} \cdot f_{PUSHOUT}$$
 (6.32)

in which  $R_{Bi0}$  is the base resistance at  $v_{BE}=0V$ . The factors  $f_{BWM}$ ,  $f_{CM}$ ,  $f_{CROWDING}$  and  $f_{PUSHOUT}$  vary with between 1 and 0, and they model the influence of base width modulation, conductivity modulation, current crowding and pushout, respectively. When the bias current increase, these factors decrease. As mentioned above, the factor  $f_{CROWDING}$  depends on the other factors.

# 6.4.2 DC and AC base resistance and the nonlinearity coefficients

In contrast with the implementation of the base resistance in most SPICE simulators [Anto 88, Hspi 96], in [Taft 91, Chiu 92] a distinction is made between the DC base resistance and the AC base resistance. An expression for the DC and AC resistance is now derived. The results of this derivation are not restricted for use with the models of [Taft 91, Chiu 92] only. In fact, they can be used for the power model as well. Although the values obtained with the power model are not exact, as stated in the previous section, they can be used for an estimation of the AC base resistance and its higher-order derivatives. This way of working can be used if the only information about the technology is a set of SPICE parameters, which are derived for the power model

After a derivation of the AC and DC base resistance, these resistances will be evaluated both for the power model and the model of Chiu. The values for the parameters of the latter model have been taken from the original paper [Chiu 92].

The base resistance is split into an intrinsic part  $R_{Bi}$  and an extrinsic part  $R_{Bex}$ :

$$R_B = R_{Bi} + R_{Bex} \tag{6.33}$$

The AC value and the DC value of the extrinsic part is assumed to be the same, since it is assumed that this part is bias independent:

$$R_{Bex} = r_{Bex} \tag{6.34}$$

In the SPICE model for the base resistance, this extrinsic part is denoted by the model parameter RBM. In the rest of the derivation,  $R_{Bex}$  is neglected. This is just done for the ease of the derivation: the results remain valid when  $R_{Bex}$  is taken into account.  $R_{Bex}$  only has to be placed in series with the AC or DC intrinsic base resistance. Further, the emitter resistance is neglected as well. With these simplifications,  $v_{B'E}$  is given by  $v_{B'B} + v_{BE}$ , in which  $v_{BB'}$  is the voltage drop over the intrinsic base resistance  $R_{Bi}$  and  $v_{BE}$  is the voltage drop over the diode in Figure 3.9b. Hence, the DC base resistance  $R_{Bi}$  is found from the following relationship

$$v_{B'E} = v_{BB'} + v_{BE} (6.35)$$

$$=R_{Bi}\cdot i_B+v_{BE} \tag{6.36}$$

in which  $i_B = I_S/\beta_F \exp{(v_{BE}/(n_F V_t))}$ . The relationship between the voltage drop over the intrinsic base resistance and the base current depends on the model that is used. By taking the derivative with respect to  $i_B$  of both sides of equation (6.36), the AC base resistance  $r_{Bi}$  is found to be

$$r_{Bi} = \frac{di_B R_{Bi}}{di_B} \tag{6.37}$$

$$=R_{Bi}+i_B\cdot\frac{dR_{Bi}}{di_B}\tag{6.38}$$

For the description of the nonlinearity of the base resistance, the first and second derivative of  $r_{Bi}$  with respect to  $i_B$  are required. Using algebra we find

$$\frac{dr_{Bi}}{di_B} = 2 \cdot \frac{dR_{Bi}}{di_B} + i_B \cdot \frac{d^2R_{Bi}}{di_B^2} \tag{6.39}$$

and

$$\frac{d^2r_{Bi}}{di_B^2} = 3 \cdot \frac{d^2R_{Bi}}{di_B^2} + i_B \cdot \frac{d^3R_{Bi}}{di_B^3} \tag{6.40}$$

These expressions can now be used for the determination of the second- and third-order nonlinearity coefficients of the intrinsic AC base resistance. These coefficients, denoted by  $K_{2r_{Bi}}$  and  $K_{3r_{Bi}}$ , are given by

$$K_{2r_{Bi}} = \frac{1}{2} \cdot \frac{dr_{Bi}}{di_B} \tag{6.41}$$

$$K_{3r_{Bi}} = \frac{1}{6} \cdot \frac{d^2 r_{Bi}}{di_B^2} \tag{6.42}$$

Since most nonlinearities that are discussed in this book are described as nonlinear admittances (conductances, transconductances, capacitances and multidimensional conductances), it is interesting to describe the nonlinearity of the intrinsic base resistance in terms of the nonlinearity coefficients of the equivalent conductance  $g_{Bi} = 1/r_{Bi}$ . In this way, the nonlinearity of the base resistance can be better compared to other nonlinearities which are described as admittances. Using the expressions (3.40) through (3.42) for the computation of the nonlinearity coefficients of a conductance when the resistance is given, one finds

$$K_{2g_{Bi}} = -\frac{K_{2r_{Bi}}}{r_{Bi}^3} \tag{6.43}$$

$$K_{3g_{Bi}} = \frac{1}{r_{Bi}^4} \left( -K_{3r_{Bi}} + 2 \frac{\left(K_{2r_{Bi}}\right)^2}{r_{Bi}} \right) \tag{6.44}$$

These nonlinearity coefficients will be evaluated in the next section.

#### 6.4.3 Evaluation of the nonlinearity coefficients

Having derived expressions for the DC and the AC base resistance, it is interesting to evaluate the difference between these two quantities. For this evaluation two base resistance models will be used. First, the power model will be used: this model is still the most widely used, even for modern bipolar transistor. Secondly, the model of Chiu [Chiu 92] is used. This model has been proven to be more accurate [Asti 96]. In addition, the nonlinearity coefficients  $K_{2g_{Bi}}$  and  $K_{3g_{Bi}}$  for both models will be discussed as well.

Figure 6.9 shows the DC and the AC value of the intrinsic base resistance as a function of the collector current and evaluated with the power model and the model of Chiu. The parameters of the latter model are taken from the original paper of Chiu [Chiu 92]. With these parameters the AC and DC base resistance at low currents is found to be  $565\Omega$ . Further, the transistor beta equals 60. The parameters of the power model have been fitted onto the value of the DC base resistance as a function of the base current according to Chiu's model. The only parameter of the power model that has to be determined in addition to the base resistance at low base currents is the parameter  $I_{RB}$ , which is the base current at which the base resistance has fallen down to 50% of the value at low currents. This parameter can be easily determined by fitting.

First, it is seen that for both models the intrinsic DC base resistance and the AC base resistance at low base currents are about equal. Indeed, at low currents, the base resistance does not change with the base current, and hence it can be regarded as a linear resistor. At higher currents, both the DC and AC resistance decrease, but the AC value is always lower. The reason for a difference between the AC and DC value is explained in Section 3.2.2.2. The intrinsic DC base resistance is the ratio of the voltage over the intrinsic part of the base region and the current that flows to the base contact through this region. When this voltage is evaluated at each base current, then the slope of this curve is the AC intrinsic base resistance.

At high base currents both the DC and AC resistance evaluated with the model of Chiu are lower than with the power model. This is mainly due to conductivity modulation which is only taken into account in the model of Chiu. The increase of the majority carriers above the impurity

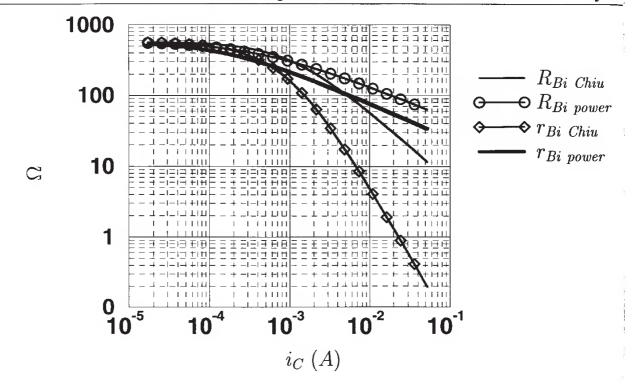


Figure 6.9: DC and AC value of the base resistance as a function of the collector with the power model and the model of Chiu.

concentration lowers the conductivity in the intrinsic base region. Also, it is seen that at high currents the ratio between the AC and DC resistance with the model of Chiu is significantly lower than with the power model. At high collector currents (or base currents) the AC intrinsic base resistance becomes so small that it is negligible compared to the extrinsic base resistance and  $r_{\pi}$ .

In order to have an idea about how strong the base resistance nonlinearity is, it is interesting to consider nonlinearity coefficients of the base conductance  $g_{Bi}=1/r_{Bi}$ . Figure 6.10 shows the second-order normalized nonlinearity coefficient  $K'_{2g_{Bi}}$  as a function of the collector current for both the power model and the model of Chiu. It is seen that at low currents the normalized coefficients computed with the power model and the model of Chiu are of the same order of magnitude. At higher currents,  $K'_{2g_{Bi}}$  with the model of Chiu increases. This is due to the fact that the conductivity of the base region increases sharply. The same conclusion can be drawn for the third-order nonlinearity coefficients. The third-order coefficient  $K_{3g_{Bi}}$  obtained with the model of Chiu is shown in Figure 6.11 as a function of the collector current.

# 6.5 Capacitors in a bipolar transistor

In the weakly nonlinear equivalent circuit of a bipolar transistor (see Figure 6.1) three capacitors are present. These capacitors are nonlinear such that they can cause distortion.

The capacitors  $C_{\mu}$  and  $C_{CS}$  correspond to junctions which in the forward active region are inversely biased. Hence, they can be modeled as in Section 3.4.

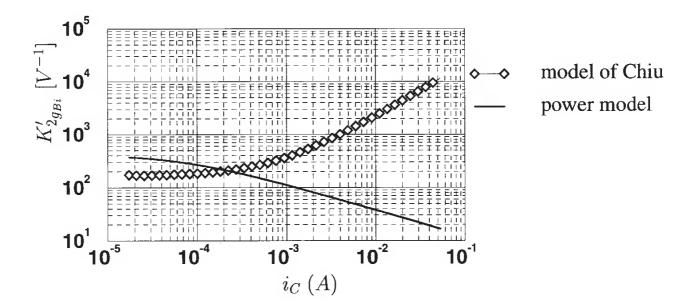


Figure 6.10: Normalized nonlinearity coefficient  $K'_{2g_{Bi}}$  with the power model and with the model of Chiu.

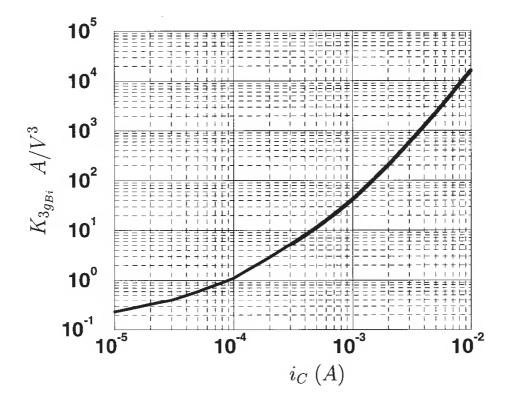


Figure 6.11: Third-order nonlinearity coefficient  $K_{3g_{Bi}}$  with the model of Chiu.

The capacitor  $C_{\pi}$  is a diffusion capacitance in parallel with a junction capacitance. The value of the junction capacitance in forward active region at sufficiently high currents is negligible

compared to the value of the diffusion capacitance. At low currents, this is not true, and the junction capacitance needs to be taken into account. In the rest of this section it will be assumed that the junction capacitance is negligible.

The diffusion capacitance originates from the variation of the excess charge of minority carriers in the base as a result of a change of  $v_{BE}$ . In Section 3.2.3 it has already been mentioned that the excess charge is given by

$$Q_D = \tau_F i_C \tag{6.45}$$

in which  $\tau_F$  is the *forward transit time*. This transit time consists of several components [Getr 76]

- 1. the delay time  $\tau_B$ , which can be interpreted as the base transit time. This is the average time in which minority carriers diffuse through the neutral base region from the emitter side to the collector side.
- 2. the delay time  $\tau_E$  associated with stored minority carrier charge in the neutral emitter region.
- 3. the emitter-base space-charge layer transit time  $\tau_{EB_{SCL}}$ .
- 4. the base-collector space-charge layer transit time  $\tau_{CB_{SCL}}$ .

The delay time  $\tau_B$  increases at high currents, due to base pushout: as the neutral base region effectively widens, it takes more time for the minority carriers to cross this region. Consequently  $\tau_F$  becomes a function of the collector current. However, in this analysis the variation of  $\tau_F$  with the collector current will be neglected. Hence, the analysis is restricted to currents below the region in which base pushout occurs. This is not a severe limitation in practice, since transistor are seldom biased in the base pushout region.

If a constant  $au_F$  is assumed, then the diffusion capacitance associated with the charge  $Q_D$  is found to be

$$C_{\pi} = \frac{dQ_D}{dv_{BE}} = \tau_F g_m \tag{6.46}$$

It is seen that  $C_{\pi}$  is directly proportional to  $g_m$ . Under low-injection conditions  $g_m = I_C/(n_F V_t)$ . Hence we find for the higher-order derivatives

$$K_{2C_{\pi}} = \frac{\tau_F g_m}{2n_F V_t} = \tau_F K_{2g_m} \tag{6.47}$$

$$K_{3C_{\pi}} = \frac{\tau_F g_m}{6n_F^2 V_t^2} = \tau_F K_{3g_m} \tag{6.48}$$

The expressions for the nonlinearity coefficients of  $C_{\pi}$  reveal that the nonlinearity of  $C_{\pi}$  tracks the nonlinearity of  $g_m$ . Hence they have the same normalized nonlinearity coefficients:

$$K_{2C_{\pi}}' = K_{2g_m}' = \frac{1}{2n_F V_t} \tag{6.49}$$

$$K_{3C_{\pi}}' = K_{3g_m}' = \frac{1}{6n_F^2 V_t^2} \tag{6.50}$$

It is interesting to compare the nonlinearity of  $C_{\pi}$  to the nonlinearity of  $C_{\mu}$  and  $C_{cs}$ . This can be done with a comparison of the normalized nonlinearity coefficients. At room temperature,  $K'_{2C_{\pi}} \approx 20V^{-1}$  whereas the second-order normalized nonlinearity coefficients  $K'_{2C_{\mu}}$  and  $K'_{2C_{CS}}$  are in the range of  $-0.1V^{-1}$  to  $-0.25V^{-1}$  as pointed out in Section 3.4. Hence, a diffusion capacitance is by far more nonlinear than a junction capacitance. This is due to the exponential dependence of the capacitance value on the voltage over the capacitance. The third-order normalized nonlinearity coefficients of junction capacitors are in the order of  $0.2V^{-2}$  and lower.

# 6.6 Summary

In this chapter we discussed the nonlinearity coefficients that describe the different basic non-linearities of a bipolar transistor. Hereby we have used the Gummel-Poon model. In addition we considered the nonlinearity of the Early effect which is not modeled in the Gummel-Poon model. For the base resistance we have considered both the classical power model and the model of Chiu that takes into account current crowding and base conductivity modulation.

# **Chapter 7**

# MOS transistor models for distortion analysis

### 7.1 Introduction

Due to continuous advances in technology the MOS transistor has scaled down rapidly. Whereas  $3\mu m$  CMOS was used in the middle of the eighties, technologies with effective gate lengths of  $0.35\mu m$  are in use in 1997. In these small devices several physical effects play a much larger role than in transistors of older technologies. Some of these effects have given rise to additional technological steps in the processing of a MOS transistor, such as the use of lightly-doped drains. As a result, a modern deep submicron transistor is a more complicated structure than a transistor of older technologies.

The behavior of MOS transistors of old technologies can be approximated well by assuming that the current flow in the transistor is one-dimensional. On the other hand, for the modelling of a modern MOS transistor several two-dimensional and even three-dimensional effects need to be taken into account.

Due to the complex behavior of a MOS transistor, many MOS transistor models have been proposed. Some of these models have been listed in the references [VdWie 79, White 80, Liu 82, Gueb 83, Maes 84, Anto 88, Gar 87a, Sheu 87, Toh 88, Yu 88, Moon 91, Hspi 96, Park 92, Pow 9 Chow 92b, Gow 91, Gow 93, BSIM 95, Hua 93, Cheng 96, Enz 95, Matt 96, Klaas 96]. Many of these models are in use. For example, commercial versions of SPICE [Hspi 96] support about fifty MOS models, some of them being private to large IC companies. Moreover, models are continuously changing as a result of scaling and existing models may become inaccurate for small devices.

In the seventies the simple quadratic model of Shichman and Hodges [Shich 68] was used to model MOS transistors in analog circuit design. This model has been implemented in SPICE simulators as the level 1 model [Anto 88, Hspi 96], was used However, this model is outdated for the accurate simulation of modern MOS transistors in deep submicron technologies. The reason is that effects that were recognized in the seventies as second-order effects have become more important for scaled devices. The modeling of these effects has required and still requires much

7.1 Introduction

201

research effort.

A computation of the nonlinearity coefficients from the model equations requires that the model equations are sufficiently accurate and that the model parameters have been extracted accurately. However, the modeling of transistors for computer-aided design has been driven by the needs of digital circuit designers for years [Tsiv 93b]. Moreover, even if a model intrinsically could do a decent job over certain bias ranges of the device, it is often not given the opportunity to do so due to a poor parameter extraction. For example, for a digital circuit designer a reasonable I-V characteristic accuracy is important, whereas a wrong slope in those characteristics is usually of no great consequence. As a result, small-signal parameters that are obtained by differentiation, are not accurate. The situation is even worse for the values of the higher-order derivatives that determine the different nonlinearity coefficients of the transistor nonlinearities.

In the analog design community efforts are done to overcome these accuracy problems at least for the small-signal parameters [Tsiv 93b, Gow 93, Enz 95, BSIM 95]. If the accuracy on the small-signal parameters is met, then the accuracy problem is shifted to the higher-order derivatives. Since a good accuracy on these derivatives is seldom important in analog circuit design, except for harmonic and intermodulation distortion, which often are still secondary specifications, little attention is paid to the accuracy of these derivatives. With the development of more analog integrated circuits for RF and telecommunications, specifications on nonlinear distortion might become more important, which would then require an accurate characterization of nonlinear distortion.

More than for the bipolar transistor, the reader should be aware that with existing MOS models that are described by analytical equations, nonlinearity coefficients can only be predicted with a limited accuracy. This will lead to errors on the prediction of harmonic or intermodulation distortion of a few decibels. If a better accuracy is required, then numerical techniques as used in device simulators will be required. One of the reasons for this lower accuracy compared to bipolar transistors is that in MOS transistors many effects, some of which are two- or three-dimensional, occur simultaneously. In order to end up with a MOS model that consists of explicit equations that can be evaluated efficiently, it is often assumed that the different effects are small enough such that they do not interact. This approach might yield a satisfactory accuracy on the current, but the error on the derivatives will tend to increase with the order of the derivative.

The emphasis in this chapter is on the quasi-static behavior of a MOS transistor in strong inversion operation, both in the triode and saturation region. Strong inversion operation means that at least one end of the channel is strongly inverted. Weak inversion is briefly discussed, whereas moderate inversion, which is the regime between weak and strong inversion is not treated in this chapter. Other "transition regions" such as the operating region around the onset of saturation, are not treated in detail. Their analytical treatment is not straightforward. Instead, it is assumed that a transistor is biased in an operating point that is sufficiently far from these transition regions and, in addition, it is assumed that the signal swings are small enough such that a transistor keeps on operating in the same region.

Since a MOS transistor is a four-terminal device, the drain current is a function of three terminal voltages. The nonlinearity coefficients that are needed to describe the three-dimensional nonlinear drain current are defined in Section 7.2.

The largest part of this chapter is devoted to the calculation of these nonlinearity coefficients.

The nonlinearities of the capacitors in a MOS transistor will only be discussed briefly.

The nonlinearity coefficients will be computed using different models although it is not the goal to present an extended comparison of many MOS models such as in the excellent work of [Foty 96]. Aspects of the SPICE level 1, 2 and 3 models and the BSIM models will be considered, while some other models that are less popular will be mentioned where it is appropriate. The nonlinearity coefficients will be evaluated for long-channel transistors as well as for short-channel transistors with gate lengths down to  $0.5\mu m$ . Two reference technologies will be used for numerical evaluations, namely a  $0.7\mu m$  and a  $0.5\mu m$  technology. The most important SPICE parameters of these technologies for the SPICE levels 1, 2 and 3 are listed in Table 7.1. The meaning of these parameters will be discussed further in this chapter.

SPICE parameter	meaning	"physical" parameter	used in level(s)	value in $0.7 \mu m$ process	value in $0.5 \mu m$ process
VTO	zero-bias gate-source extrapolated threshold voltage	$V_{TO}$	1, 2, 3	0.75V	0.57V
TOX	gate-oxide thickness	$t_{ox}$	1, 2, 3	$1.7 \times 10^{-8} m$	$10^{-8}m$
GAMMA	body-effect coefficient	γ	1, 2, 3	$0.75V^{1/2}$	$0.69V^{1/2}$
PHI	surface inversion potential	φ	1, 2, 3	0.8V	0.9V
UO	surface mobility	$\mu_0$	1, 2, 3	$0.047m^2/(V.s)$	$0.046m^2/(V.s)$
THETA	mobility-reduction coefficient	θ	3	$0.079V^{-1}$	$0.129V^{-1}$
VMAX	saturation velocity	$v_{sat}$	3	$1.94 \times 10^5 m/s$	$1.3 \times 10^5 m/s$
UCRIT	critical field for mobility reduction		2	$1.08 \times 10^7 V/m$	
UEXP	critical field exponent		2	0.124	
LAMBDA	channel-length modulation factor	λ	1,2	$0.0364V^{-1}$	

Table 7.1: SPICE parameters for level 1, 2 and 3 of an n-MOS transistor in the  $0.7\mu m$  and  $0.5\mu m$  technologies that will be used in this book.

It must be noted that some of these parameters are fit parameters and their value does not necessarily correspond to the value of the corresponding physical (or semi-empirical) parameter.

For example, the value of the SPICE parameter VMAX is not equal to the saturation velocity. This issue will be discussed further in the subsequent sections. Level 2 parameters for the  $0.5 \mu m$  transistor are not listed, since they were not available. For the  $0.5 \mu m$  process, BSIM3 model parameters are available as well. Since the list of parameters for this model is long, they are not shown in Table 7.1. Instead, values will be given when they are needed throughout this chapter.

The nonlinearity coefficients for the drain current will first be derived in Sections 7.4 and 7.5 for the strong inversion regime without taking into account any effects that are caused by scaling down the technology. Hence, a thick gate oxide is assumed, such that the influence of the normal electric field, i.e. the field perpendicular to the channel, is neglected. A neglection of the normal field is reflected in the computations by setting the parameter  $\theta$  of Table 7.1 to zero, or, for the level 2 model, setting UEXP to zero. Further, velocity saturation will be neglected initially. Later in this chapter, in Sections 7.6 and 7.7 the influence of the normal field and velocity saturation on the nonlinearity coefficients will be considered in detail. Next, the following effects will be considered in Sections 7.8, 7.9, 7.10 and 7.11: nonuniform doping effects, dependence of the threshold voltage on the channel length (drain induced barrier lowering) and width of the transistor, the influence of source and drain resistances, channel-length modulation, drain-induced barrier lowering and the substrate current. This list does not cover all effects that occur in a MOS transistor. For example, the poly-gate depletion effect [BSIM 95] is not considered.

Section 7.12 briefly covers the nonlinearity of capacitors in a MOS transistor in strong inversion. Finally, the nonlinearity of the drain current in weak inversion is briefly discussed in Section 7.13.

Before the current equations are presented, it is instructive to see how an expression for the drain current is derived in general. This will be explained in Section 7.3. In this way the reader will have a better insight in the simplifications made in many models. This is important, since some simplifications may be too rough such that the error on some nonlinearity coefficients is very large.

The analyses in this chapter will be performed on n-MOS transistors. The analysis for p-MOS transistors is similar. The same expressions can be used if absolute values for the terminal voltages are considered. Also, absolute values must be considered for the threshold voltage  $V_{TO}$  and for the Fermi potential.

Finally, it must be remarked that the symbols that are used in this chapter are explained in the list of symbols at the beginning of this book.

# 7.2 Nonlinearity coefficients of the three-dimensional drain current nonlinearity

A MOS transistor is a four-terminal device, the terminals being gate, drain, source and bulk. The voltages in the transistor are either referred to the source, which is done for example in the SPICE models level 1, 2 and 3 [Anto 88, Hspi 96], or to the bulk, as is done for example in [Enz 95] and partially in [Tsiv 88]. We will consider both reference systems in this chapter.

The nonlinear equivalent circuit of a MOS transistor is depicted in Figure 7.1. This equiv-

alent circuit is valid for low and medium-high frequencies, typically up to a tenth of the cutoff frequency [Tsiv 88]. MOS models that are valid up to higher frequencies are discussed for example in [Tsiv 88, Park 92, Enz 95, BSIM 95].

Three nonlinear elements occur in the equivalent circuit of Figure 7.1:

- the capacitors  $C_{sb}$  and  $C_{db}$  between source and bulk and drain and bulk. These will be discussed in Section 7.12.
- the drain current  $i_D$ . This current is a function of three voltages. When the voltages are referred to the source, then the current is a function of the voltages  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ . When all voltages are referred to the bulk, then the current is considered as a function of  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ .

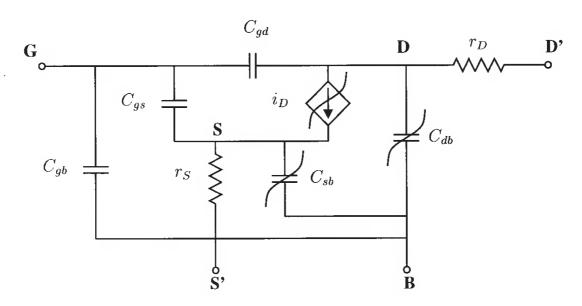


Figure 7.1: Nonlinear model of a MOS transistor.

The other elements in the MOS equivalent circuit are most often considered as linear elements. These are the source and drain resistances  $^{1}$   $r_{D}$  and  $r_{S}$  and the capacitors between the gate and the other terminals, namely  $C_{gs}$ ,  $C_{gd}$  and  $C_{gb}$ . The resistors  $r_{D}$  and  $r_{S}$  will be neglected initially, until their influence will be discussed in Section 7.10. Although we consider the oxide capacitors as linear elements in Figure 7.1, we will see that this is only a good approximation when the transistor is in the saturation region. We will see that in the triode region the oxide capacitors are bias dependent.

For very small signals, the circuit of Figure 7.1 can be linearized. The linearized equivalent circuit of Figure 7.1 is shown in Figure 7.2. Hereby, it is assumed that the voltages are referred to the source. Then one defines

$$g_m = \frac{\partial i_D}{\partial v_{GS}}$$
  $g_o = \frac{\partial i_D}{\partial v_{DS}}$   $g_{mb} = -\frac{\partial i_D}{\partial v_{SB}}$  (7.1)

<sup>&</sup>lt;sup>1</sup>For sub-micron devices the gate resistance will become more and more important, especially at high frequencies [Klaas 96]. This means that an extra circuit element needs to be added to the schematic of Figure 7.1.

These coefficients define three components of the drain current, each of them being linearly proportional to one of the three controlling voltages of the drain current.

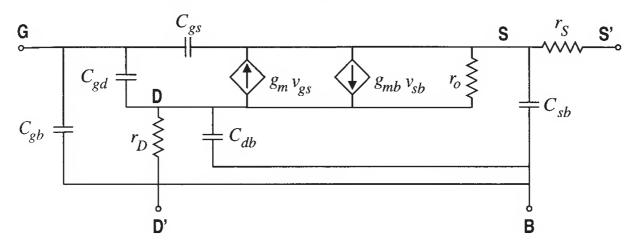


Figure 7.2: Small-signal model of a MOS transistor with the source as a reference.

Figure 7.3 shows the small-signal equivalent circuit of a MOS transistor when the voltages are referred to the bulk. The transconductances in this circuit are defined as

$$g_{mg} = \frac{\partial i_D}{\partial v_{GB}} \quad g_{md} = \frac{\partial i_D}{\partial v_{DB}} \quad g_{ms} = -\frac{\partial i_D}{\partial v_{SB}}$$
 (7.2)

Here the same symbols are used as the ones defined in [Enz 95]. The small-signal parameters

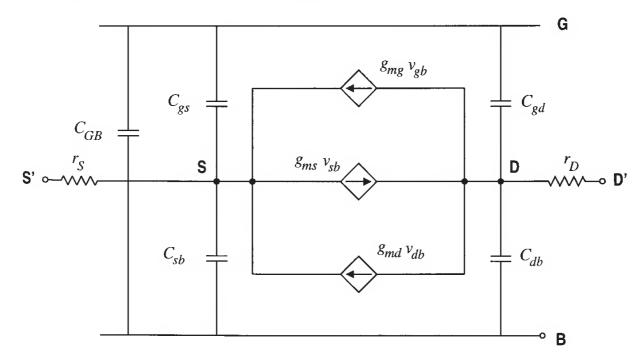


Figure 7.3: Small-signal model of a MOS transistor with the bulk as a reference.

defined in equations (7.1) and (7.2) are nothing else but the first-order terms of the power series description of the drain current as a function of three variables, in the same way as described in Section 3.2.5.

When signals of larger amplitude are applied to the MOS transistor, then the linear approximation no longer holds and we have to describe the drain current with a nonlinear function of three voltages. This function can be expanded into a three-dimensional power series, as we did in Section 3.2.5 for a general three-dimensional conductance. Depending on the reference terminal that is used (source or bulk), different nonlinearity coefficients will be used.

#### 7.2.1 Coefficients referred to the source

When the drain current is described as a function of  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ , then the AC drain current can be described by the following three-dimensional power series

$$i_{d} = g_{m} \cdot v_{gs} + K_{2g_{m}} \cdot v_{gs}^{2} + K_{3g_{m}} \cdot v_{gs}^{3} + \dots + g_{o} \cdot v_{ds} + K_{2g_{o}} \cdot v_{ds}^{2} + K_{3g_{o}} \cdot v_{ds}^{3} + \dots - g_{mb} \cdot v_{sb} - K_{2g_{mb}} \cdot v_{sb}^{2} - K_{3g_{mb}} \cdot v_{sb}^{3} + \dots + K_{2g_{m} \& g_{mb}} \cdot v_{gs} \cdot v_{sb} + K_{3g_{m} \& g_{mb}} \cdot v_{gs}^{2} \cdot v_{sb} + K_{3g_{m} \& 2g_{mb}} \cdot v_{gs} \cdot v_{sb}^{2} + \dots + K_{2g_{m} \& g_{o}} \cdot v_{gs} \cdot v_{ds} + K_{3g_{m} \& g_{o}} \cdot v_{gs}^{2} \cdot v_{ds} + K_{3g_{m} \& 2g_{o}} \cdot v_{gs} \cdot v_{ds}^{2} + \dots + K_{2g_{mb} \& g_{o}} \cdot v_{sb} \cdot v_{ds} + K_{3g_{mb} \& g_{o}} \cdot v_{sb}^{2} \cdot v_{ds} + K_{3g_{mb} \& 2g_{o}} \cdot v_{sb} \cdot v_{ds}^{2} + \dots + K_{3g_{m} \& g_{mb} \& g_{o}} \cdot v_{gs} \cdot v_{sb} \cdot v_{ds} + \dots$$

$$+ K_{3g_{m} \& g_{mb} \& g_{o}} \cdot v_{gs} \cdot v_{sb} \cdot v_{ds} + \dots$$

$$+ K_{3g_{m} \& g_{mb} \& g_{o}} \cdot v_{gs} \cdot v_{sb} \cdot v_{ds} + \dots$$

It should be remarked once again that the terms of the powers of  $v_{sb}$  only are negative. The reason is that the bulk transconductance  $g_{mb}$  is usually represented as a controlled source flowing from the source to the drain, which is opposite to the direction of the source  $g_m v_{gs}$ . This yields a positive value for  $g_{mb}$ . Hence,  $g_{mb}$  is the first-order derivative times -1. For consistency, the coefficients  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  are adjusted in the same way.

The notation of the coefficients that occur in the series of equation (7.3) corresponds to the convention made in Section 3.2.5. For clarity, the definitions of the nonlinearity coefficients are resumed in Table 7.2.

# 7.2.2 Coefficients referred to the bulk

When voltages are referred to the bulk, then the power series for the AC drain current is given by

$$\begin{split} i_d &= g_{mg} \cdot v_{gb} + K_{2g_{mg}} \cdot v_{gb}^2 + K_{3g_{mg}} \cdot v_{gb}^3 + \dots \\ &+ g_{md} \cdot v_{db} + K_{2g_{md}} \cdot v_{db}^2 + K_{3g_{md}} \cdot v_{db}^3 + \dots \\ &- g_{ms} \cdot v_{sb} - K_{2g_{ms}} \cdot v_{sb}^2 - K_{3g_{ms}} \cdot v_{sb}^3 + \dots \end{split}$$

	ı		
$g_m$	$rac{\partial i_D}{\partial v_{GS}}$	$K_{3_{2g_m}\&g_o}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GS}^2 \partial v_{DS}}$
$K_{2g_m}$	$\frac{1}{2} \frac{\partial^2 i_D}{\partial v_{GS}^2}$	$K_{3_{g_m\&2g_o}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GS} \partial v_{DS}^2}$
$K_{3g_m}$	$\boxed{\frac{1}{6} \frac{\partial^3 i_D}{\partial v_{GS}^3}}$	$K_{2_{g_m}\&g_{mb}}$	$rac{\partial^2 i_D}{\partial v_{GS} \partial v_{SB}}$
$g_{mb}$	$-rac{\partial i_D}{\partial v_{SB}}$	$K_{3_{2g_m}\&g_{mb}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GS}^2 \partial v_{SB}}$
$K_{2g_{mb}}$	$-\frac{1}{2}\frac{\partial^2 i_D}{\partial v_{SB}^2}$	$K_{3_{g_m}\&2g_{mb}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GS} \partial v_{SB}^2}$
$K_{3g_{mb}}$	$-\frac{1}{6}\frac{\partial^3 i_D}{\partial v_{SB}^3}$	$K_{2_{g_{mb}}\&g_o}$	$\frac{\partial^2 i_D}{\partial v_{SB} \partial v_{DS}}$
$g_o$	$rac{\partial i_D}{\partial v_{DS}}$	$K_{3_{2g_{mb}\&g_o}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{SB}^2 \partial v_{DS}}$
$K_{2g_o}$	$\frac{1}{2} \frac{\partial^2 i_D}{\partial v_{DS}^2}$	$K_{3_{g_{mb}\&2g_o}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{SB} \partial v_{DS}^2}$
$K_{3g_o}$	$\boxed{\frac{1}{6} \frac{\partial^3 i_D}{\partial v_{DS}^3}}$	$K_{3_{g_m\&g_{mb}\&g_o}}$	$\frac{\partial^3 i_D}{\partial v_{GS} \partial v_{SB} \partial v_{DS}}$
$K_{2_{g_m}\&g_o}$	$\frac{\partial^2 i_D}{\partial v_{GS} \partial v_{DS}}$		

Table 7.2: Definition of the nonlinearity coefficients for the AC drain current as a function of the terminal voltages referred to the source.

$$+ K_{2g_{mg}\&g_{ms}} \cdot v_{gb} \cdot v_{sb} + K_{32g_{mg}\&g_{ms}} \cdot v_{gb}^{2} \cdot v_{sb} + K_{3g_{mg}\&2g_{ms}} \cdot v_{gb} \cdot v_{sb}^{2} + \dots$$

$$+ K_{2g_{mg}\&g_{md}} \cdot v_{gb} \cdot v_{db} + K_{32g_{mg}\&g_{md}} \cdot v_{gb}^{2} \cdot v_{db} + K_{3g_{mg}\&2g_{md}} \cdot v_{gb} \cdot v_{db}^{2} + \dots$$

$$+ K_{2g_{ms}\&g_{md}} \cdot v_{sb} \cdot v_{db} + K_{32g_{ms}\&g_{md}} \cdot v_{sb}^{2} \cdot v_{db} + K_{3g_{ms}\&2g_{md}} \cdot v_{sb} \cdot v_{db}^{2} + \dots$$

$$+ K_{3g_{mg}\&g_{ms}\&g_{md}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db} + \dots$$

$$(7.4)$$

The definitions of the nonlinearity coefficients used in this equation are given in Table 7.3. In

$g_{mg}$	$rac{\partial i_D}{\partial v_{GB}}$	$K_{3_{2g_{mg}\&g_{md}}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB}^2 \partial v_{DB}}$
$K_{2g_{mg}}$	$\frac{1}{2} \frac{\partial^2 i_D}{\partial v_{GB}^2}$	$K_{3_{g_{mg}\&2g_{md}}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB} \partial v_{DB}^2}$
$K_{3g_{mg}}$	$\frac{1}{6} \frac{\partial^3 i_D}{\partial v_{GB}^3}$	$K_{2_{g_{mg}\&g_{ms}}}$	$rac{\partial^2 i_D}{\partial v_{GB}\partial v_{SB}}$
$g_{ms}$	$-rac{\partial i_D}{\partial v_{SB}}$	$K_{3_{2g_{mg}\&g_{ms}}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB}^2 \partial v_{SB}}$
$K_{2g_{ms}}$	$-\frac{1}{2}\frac{\partial^2 i_D}{\partial v_{SB}^2}$	$K_{3_{g_{mg}\&2g_{ms}}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB} \partial v_{SB}^2}$
$K_{3g_{ms}}$	$-\frac{1}{6}\frac{\partial^3 i_D}{\partial v_{SB}^3}$	$K_{2_{g_{ms}}\&g_{md}}$	$rac{\partial^2 i_D}{\partial v_{SB} \partial v_{DB}}$
$g_{md}$	$rac{\partial i_D}{\partial v_{DB}}$	$K_{3_{2g_{ms}\&g_{md}}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{SB}^2 \partial v_{DB}}$
$K_{2g_{md}}$	$\boxed{\frac{1}{2} \frac{\partial^2 i_D}{\partial v_{DB}^2}}$	$K_{3_{g_{ms}}\&2g_{md}}$	$\frac{1}{2} \frac{\partial^3 i_D}{\partial v_{SB} \partial v_{DB}^2}$
$K_{3g_{md}}$	$\frac{1}{6} \frac{\partial^3 i_D}{\partial v_{DB}^3}$	$K_{3_{g_{mg}}\&g_{ms}\&g_{md}}$	$\frac{\partial^3 i_D}{\partial v_{GB} \partial v_{SB} \partial v_{DB}}$
$K_{2_{g_{mg}}\&g_{md}}$	$\frac{\partial^2 i_D}{\partial v_{GB} \partial v_{DB}}$		

Table 7.3: Definition of the nonlinearity coefficients for the AC drain current as a function of the terminal voltages referred to the bulk.

Figure 7.3 it was seen that the orientation of the current source controlled by  $v_{SB}$  differs from the orientation of the other two controlled sources. This explains the three negative terms in the series expansion of equation (7.4). This is similar to the different sign of the terms that are proportional to the derivatives with respect to  $v_{SB}$  in equation (7.3).

# 7.2.3 Relationship between the coefficients of the two reference systems

The coefficients defined in the Sections 7.2.1 and 7.2.2 are not independent. This is clear since they describe the same drain current but in terms of derivatives to other voltages. In Appendix E the relationship between the two sets of coefficients are computed. The main results of these computations are given here.

First, it is found that nonlinearity coefficients that are proportional to a derivative with respect to  $v_{GS}$  only, are identical to nonlinearity coefficients that are proportional to a derivative with respect to  $v_{GB}$  only. This means that

$$g_{mg} = g_m \tag{7.5}$$

$$K_{2q_{mq}} = K_{2q_m} (7.6)$$

$$K_{3g_{mg}} = K_{3g_m} (7.7)$$

A similar identity is found for derivatives with respect to  $v_{DS}$  and  $v_{DB}$ . This yields

$$g_{md} = g_o (7.8)$$

$$K_{2q_{md}} = K_{2q_0} (7.9)$$

$$K_{3g_{md}} = K_{3g_o} (7.10)$$

The relationships between  $g_{mb}$  and  $g_{ms}$  and the corresponding higher-order derivatives are somewhat more complicated. One finds

$$g_{mb} = -g_{mq} - g_{md} + g_{ms} (7.11)$$

At  $v_{DS}=0V$  we will see that  $g_{mg}$  is zero. Since at  $v_{DS}=0V$  the role of source and drain is identical, it is clear that derivatives of the drain current with respect to  $v_{DB}$  are identical with derivatives with respect to  $v_{SB}$ . This means that  $g_{md}$  and  $g_{ms}$  are equal. Using to equation (7.11) we find that  $g_{mb}$  is zero. However, we will see that there are many MOS models for which the role of source and drain is not identical at  $v_{DS}=0V$ . This is not correct, and such models must be handled with care when a transistor needs to be modeled accurately around  $v_{DS}=0V$ .

The expression of  $K_{2g_{mb}}$  in terms of nonlinearity coefficients that are referred to the bulk, is given by

$$K_{2g_{mb}} = -K_{2g_{mg}} - K_{2g_{mg} \& g_{md}} - K_{2g_{mg} \& g_{ms}} - K_{2g_{md}} - K_{2g_{ms} \& g_{md}} + K_{2g_{ms}}$$

$$(7.12)$$

and for  $K_{3g_{mb}}$ 

$$K_{3g_{mb}} = -K_{3g_{mg}} - K_{3g_{mg}\&g_{md}} - K_{3g_{mg}\&g_{ms}} - K_{3g_{mg}\&2g_{md}} - K_{3g_{mg}\&2g_{ms}} - K_{3g_{md}} - K_{3g_{md}\&g_{ms}} - K_{3g_{md}\&2g_{ms}} + K_{3g_{ms}} - K_{3g_{md}\&g_{md}\&g_{ms}}$$

$$(7.13)$$

For nonlinearity coefficients that are proportional to cross-derivatives, similar relationships can be derived.

In later sections, when nonlinearity coefficients will be computed with both reference systems, the reader can verify that the above relationships hold, apart from the roundoff errors. In addition, the above relationships can be used to find the nonlinearity coefficients with a given reference system from the nonlinearity coefficients with the other reference system.

# 7.3 Basic relations for the drain current in strong inversion

Figure 7.4 depicts an n-MOS transistor in strong inversion. It is assumed that  $v_{GS}>0$ ,  $v_{SB}\geq0$ 

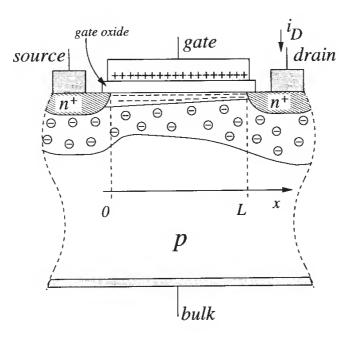


Figure 7.4: Cross section of an n-MOS transistor in strong inversion.

and  $v_{DS} \geq 0$ . When the gate potential is sufficiently high compared to the potential of source and drain, an *inversion layer* or *channel* is formed right below the gate oxide. In the *triode region* of operation, this inversion layer stretches from the source to the drain and it has a negative charge. The length of the channel is L and the width (perpendicular to the figure) is W. It is assumed that the inversion layer is of infinitesimal thickness. This approximation is referred to as the charge-sheet model [Tsiv 88]. Below the inversion layer a depletion layer is present which consists of fixed negative ions. When  $v_{DS} > 0$  then a drain current  $i_D$  flows from the drain to the source. It is assumed that the current in the channel is caused by drift of the electrons into the x-direction and not by diffusion of electrons. Under these assumptions, the drain current  $i_D$  at point x in the channel is found as [Sodi 84]

$$i_D(x) = W(-Q_I'(x)) v(x)$$
 (7.14)

in which  $(-Q_I'(x))$  is the negative inversion layer charge per unit area and v(x) is the velocity or **drift velocity** of the carriers. This velocity is assumed to be proportional to the electric field in the x-direction, also referred to as the **lateral** or **longitudinal field**  $E_x$ . The constant of proportionality between velocity and electric field is the **surface mobility**  $\mu$ :

$$v(x) = \mu E_x(x) \tag{7.15}$$

The surface mobility is nothing else but the mobility of electrons in the inversion layer right beneath the interface between silicon and the gate oxide.

The lateral field at position x is given by

$$E_x(x) = \frac{dv_{CB}}{dx} \tag{7.16}$$

in which  $v_{CB}$  is the voltage difference between point x in the channel and the bulk.

In this book we only consider the MOS transistor in a steady-state situation. In this situation, the current is constant along the channel such that we can omit the argument x from  $i_D(x)$ . Then we find from equation (7.14):

$$i_D dx = W\mu \left(-Q_I'\right) dv_{CB} \tag{7.17}$$

Integrating from x = 0 to x = L yields

$$\int_{0}^{L} i_{D} dx = W \int_{v_{SB}}^{v_{DB}} \mu \left( -Q_{I}' \right) dv_{CB}$$
 (7.18)

or

$$i_D = \frac{W}{L} \int_{v_{SB}}^{v_{DB}} \mu(-Q_I') \, dv_{CB} \tag{7.19}$$

This integral equation forms the basis for a drain current model in the triode region. The analytical computation of this integral depends on the different physical effects that are taken into account in a MOS model.

The surface mobility  $\mu$  can vary along the channel and it can depend on bias voltages. This will be discussed in Section 7.6. For the most simple models, however, it is regarded as a constant. It is set equal to the constant  $\mu_0$ , which is in fact a fit parameter. Its value is roughly equal to about half of the mobility in the bulk [Tsiv 88]. If the mobility is constant, it can be placed outside the integral.

The inversion layer charge per unit area is given by [Tsiv 88]

$$Q'_{I} = -C'_{ox} \left( v_{GB} - V_{FB} - \phi - v_{CB} - \gamma \sqrt{\phi + v_{CB}} \right)$$
 (7.20)

In this equation,  $V_{FB}$  is the flat-band voltage [Tsiv 88, Lak 94]. For the  $0.7\mu m$  process, the parameters of which have been given in Table 7.1,  $V_{FB} = -0.72V$ .

The parameter  $\phi$ , sometimes referred to as the *surface inversion potential* can be obtained from measurements. It is roughly equal to  $|2\Phi_F| + 6kT/q$  [Tsiv 88],  $\Phi_F$  being the *Fermi potential* [Sze 85, Tsiv 88, Lak 94]:

$$\Phi_F = V_t \ln \frac{N_A}{n_i} \tag{7.21}$$

in which  $N_A$  is the concentration of acceptors in the bulk and  $n_i$  is the intrinsic carrier concentration. At room temperature (300 degrees Kelvin) the value of  $n_i$  is approximately  $1.45 \times 10^{10} cm^{-3}$ .

Further, the parameter  $\gamma$  is the so-called body-effect coefficient [Tsiv 88, Lak 94]. This is given by

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_A}}{C'_{ox}} \tag{7.22}$$

The symbol  $C'_{ox}$  is the gate oxide capacitance per unit area:

$$C'_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{7.23}$$

Here  $t_{ox}$  is the gate-oxide thickness and  $\varepsilon_{ox}$  is the dielectric permittivity of silicon, which is about 34.5pF/m. The last term in equation (7.20) is the charge per unit area of the depletion layer underneath the inversion layer:

$$Q_B' = -\gamma C_{ox}' \sqrt{\phi + v_{CB}} \tag{7.24}$$

This charge is not constant along the channel, since it depends on  $v_{CB}$ , the voltage difference between a point in the channel and the bulk.

The above equations will serve as the basis for the derivation of the expression of the drain current with the inclusion or neglection of different effects, as discussed in the next sections.

# 7.4 Drain current in the triode region without small-geometry effects

The expression for the current in the triode region is obtained by computing the integral in equation (7.19). When small-geometry effects are neglected, the mobility can be considered as a constant such that it can be put outside of the integral.

For hand calculations and in engineering models for the drain current [Toh 88], the variation of the thickness of the depletion layer along the channel is neglected. Instead, this charge is considered as a constant, and its value is taken equal to the depletion layer charge at the source end. The current equation will first be derived for this case, and the consequences for the nonlinearity coefficients will be discussed. Next, in Section 7.4.2 the variation of the depletion layer along the channel will be taken into account.

# 7.4.1 Uniform depletion layer

As mentioned above, the thickness of the depletion layer is often considered as a constant along the channel and taken equal to the thickness at the source side of the channel. In this case, it is also appropriate to refer the voltages to the source. The integral in equation (7.19) yields

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left( (v_{GS} - V_T) v_{DS} - \frac{v_{DS}^2}{2} \right)$$
 (7.25)

The parameter  $V_T$  is the gate-source extrapolated threshold voltage [Tsiv 88] or briefly the threshold voltage, given by

$$V_T = V_{FB} + \phi + \gamma \sqrt{\phi + v_{SB}} \tag{7.26}$$

$$=V_{T0} + \gamma \left(\sqrt{\phi + v_{SB}} - \sqrt{\phi}\right) \tag{7.27}$$

with

$$V_{T0} = V_{FB} + \phi + \gamma \sqrt{\phi} \tag{7.28}$$

Equation (7.25) is widely used for hand calculations, even for small-geometry transistors. The reason is that, despite its low accuracy, it is very simple compared to more accurate models. Equation (7.25) is implemented in the *level 1 MOS model of SPICE* [Anto 88, Hspi 96]. In SPICE, the expression for the current as given in equation (7.25) is multiplied with a factor  $(1 + LAMBDA \cdot v_{DS})$ , LAMBDA being the model parameter that takes into account channel-length modulation in the saturation region (see below). This makes no physical sense, it has merely been done to obtain continuity in the transition from the triode region to saturation.

Using equation (7.25) the nonlinearity coefficients defined in equation (7.3) can now be derived. Their expression is given in Table 7.4 together with the numerical values at  $V_{DS}=0V$  and  $V_{DS}=0.45V$  for a n-MOS transistor with  $W=320\mu m$  and  $L=14\mu m$ . For both values of  $V_{DS}$  we have taken  $V_{GS}=1.9V$  and  $V_{SB}=1.3V$ . The model parameters used for the evaluation of the expressions are taken from Table 7.1 for the  $0.7\mu m$  process. The drain current at  $V_{DS}=0V$  is zero, while for  $V_{DS}=0.45V$  the current is 0.5mA.

nonlinearity coefficient	expression	Evaluation at $V_{GS} = 1.9V$ $V_{DS} = 0V$ $V_{SB} = 1.3V$	Evaluation at $V_{GS} = 1.9V$ $V_{DS} = 0.45V$ $V_{SB} = 1.3V$
$g_m$	$\frac{\mu_0 C'_{ox} W V_{DS}}{L}$	0A/V	0.982mA/V
$K_{2g_m}$	0	$0A/V^2$	$0A/V^2$
$K_{3g_m}$	0	$0A/V^3$	$0A/V^3$
$g_{mb}$	$\frac{1}{2} \frac{\mu_0 C'_{ox} W}{L} \frac{\gamma V_{DS}}{\sqrt{\phi + V_{SB}}}$	0A/V	0.254mA/V
$K_{2g_{mb}}$	$-\frac{1}{8} \frac{\mu_0 C'_{ox} W}{L} \frac{\gamma V_{DS}}{(\phi + V_{SB})^{3/2}}$	$0A/V^2$	$-30.2\mu A/V^{2}$
$K_{3g_{mb}}$	$\frac{1}{16} \frac{\mu_0 C'_{ox} W}{L} \frac{\gamma V_{DS}}{\left(\phi + V_{SB}\right)^{5/2}}$	$0A/V^3$	$7.17 \mu A/V^3$

$g_o$	$\frac{\mu_0 C'_{ox} W}{L} \left( V_{GS} - V_T - V_{DS} \right)$	1.60mA/V	0.619mA/V
$K_{2g_o}$	$-\frac{\mu_0 C'_{ox} W}{2L}$	$-1.09mA/V^2$	$-1.09mA/V^{2}$
$K_{3g_o}$	0	$0A/V^3$	$0A/V^3$
$K_{2_{g_m\&g_o}}$	$\frac{\mu_0 C'_{ox} W}{L}$	$2.18mA/V^2$	$2.18mA/V^2$
$K_{3_{2g_m\&g_o}}$	0	$0A/V^3$	$0A/V^3$
$K_{3_{g_m\&2g_o}}$	0	$0A/V^3$	$0A/V^3$
$K_{2_{g_m\&g_{mb}}}$	0	$0A/V^2$	$0A/V^2$
$K_{3_{2g_m\&g_{mb}}}$	0	$0A/V^3$	$0A/V^3$
$K_{3_{g_m\&2g_{mb}}}$	0	$0A/V^3$	$0A/V^3$
$K_{2_{g_{mb}\&g_o}}$	$-\frac{1}{2}\frac{\mu_0 C'_{ox} W}{L} \frac{\gamma}{\sqrt{\phi + V_{SB}}}$	$-0.565mA/V^2$	$-0.565mA/V^2$
$K_{3_{2g_{mb}\&g_o}}$	$\frac{1}{8} \frac{\mu_0 C'_{ox} W}{L} \frac{\gamma}{\left(\phi + V_{SB}\right)^{3/2}}$	$67.3\mu A/V^3$	$67.3\mu A/V^3$
$K_{3_{g_{mb}\&2g_o}}$	0	$0A/V^3$	$0A/V^3$
$K_{3_{g_m\&g_{mb}\&g_o}}$	0	$0A/V^3$	$0A/V^3$

Table 7.4: Nonlinearity coefficients (according to equation (7.3)) of the drain current in the triode region using the model of equation (7.25). Numerical values are obtained for a transistor with  $W=320\mu m$  and  $L=14\mu m$  with the model parameters of the  $0.7\mu m$  process (see Table 7.1).

It should be mentioned once again that the values of this table are not realistic for smal

geometries.

Some interesting conclusions can be drawn from Table 7.4. First, it is seen that  $g_m$  is constant with respect to  $v_{GS}$ . This is due to the linear dependence of  $v_{GS}$  on the drain current in the triode region. As a result,  $K_{2g_m}$  and  $K_{3g_m}$  are zero. Also, it is seen that at  $V_{DS} = 0V$ ,  $g_m$  is zero, and it increases linearly with  $v_{DS}$ .

Next, it is seen that the output conductance decreases as  $V_{DS}$  increases. At  $V_{DS} = 0V$ , its value is  $\mu_0 C'_{ox} W \left( V_{GS} - V_T \right) / L$ . It is zero when  $V_{DS} = V_{GS} - V_T$ . This corresponds to the point where the saturation region starts. This is discussed in the next section. Since  $g_o$  changes with  $V_{DS}$ , the derivative of  $g_O$  with respect to  $v_{DS}$  is not zero, and hence  $K_{2g_o}$  is not zero. The normalized second-order nonlinearity coefficient  $K'_{2g_o}$  is found to be

$$K'_{2g_o} = -\frac{1}{2\left(V_{GS} - V_T - V_{DS}\right)} \tag{7.29}$$

which, for this example, equals  $0.682V^{-1}$  at  $V_{DS}=0V$ . It is seen that this normalized nonlinearity coefficient is minimal at  $V_{DS}=0V$ . Further it is seen that the third-order nonlinearity coefficient  $K_{3q_o}$  is zero.

Next, consider the nonlinear dependence of the drain current on  $v_{SB}$ . The drain current depends on  $V_{SB}$  through  $V_T$ . The coefficients  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  are proportional to  $V_{DS}$ . Hence for zero  $V_{DS}$  they are zero. Further, it is seen that even if  $V_{SB}=0V$ ,  $g_{mb}$  and its nonlinearity coefficients are not zero. The normalized second- and third-order nonlinearity coefficients  $K'_{2g_{mb}}$  and  $K'_{3g_{mb}}$  are given by

$$K_{2g_{mb}}' = -\frac{1}{4} \frac{1}{(\phi + V_{SB})} \tag{7.30}$$

$$K_{3g_{mb}}' = \frac{1}{8} \frac{1}{(\phi + V_{SB})^2} \tag{7.31}$$

It is seen that these normalized coefficients only depend on  $V_{SB}$  and not on other voltages. Also, they decrease as  $V_{SB}$  increases. This can be seen as follows. The dependence of the current or  $V_T$  on  $V_{SB}$  is caused by the depletion layer underneath the inversion layer. The thickness of this layer depends on  $V_{SB}$ . As  $V_{SB}$  increases, the thickness of the depletion layer increases as well, but the change is less pronounced at high values of  $V_{SB}$ . This effect has also been seen with the nonlinearity coefficients of a junction capacitor (see Section 3.4), which is also determined by a depletion layer.

Next, consider the coefficients that are determined by cross-derivatives. Due to the linear dependence of the drain current with respect to  $v_{GS}$ , the cross-derivatives in which more than one differentiation is performed with respect to  $v_{GS}$ , are zero. Further, as  $g_m$  is independent of  $v_{SB}$ ,  $K_{2g}$ ,  $K_{3g}$ ,  $K_{3g}$ , and  $K_{3g}$ , are zero as well.

 $v_{SB}, K_{2_{g_m\&g_{mb}}}, K_{3_{g_m\&2g_{mb}}}$  and  $K_{3_{g_m\&g_{mb}\&g_o}}$  are zero as well. Since  $g_m$  is proportional to  $v_{DS}$ , the nonlinearity coefficient  $K_{2_{g_m\&g_o}}$  is not zero. This offers an interesting application as will be shown in the next section.

#### 7.4.1.1 Application: a single-transistor mixer

Referring to the power series of the drain current as a function of the transistor's terminal voltages (equation (7.3)), this series contains a term  $K_{2g_m\&g_o}v_{gs}v_{ds}$ . Assume now that  $v_{gs}$  and  $v_{ds}$  are sine waves:

$$v_{qs} = A_1 \sin\left(\omega_1 t\right) \tag{7.32}$$

$$v_{ds} = A_2 \sin\left(\omega_2 t\right) \tag{7.33}$$

When these values are substituted into the term  $K_{2g_m\&g_o}v_{gs}v_{ds}$  then it is seen that the drain current contains a frequency component at the sum frequency  $\omega_1+\omega_2$  and at the difference frequency  $|\omega_1-\omega_2|$ . The amplitude of these two components is given by

component at 
$$|\omega_1 \pm \omega_2| = A_1 A_2 \frac{K_{2g_m \& g_o}}{2}$$
 (7.34)

These components are the desired components for a frequency translation performed by a mixer. Hence, a MOS transistor in the triode region can be used as a mixer. This mixer belongs to the category that we termed as weakly nonlinear mixers in Chapter 2. One input signal is applied between the gate and the source, the other signal between drain and source. When the transistor is biased with  $V_{DS}=0V$ , then the amplitude of the component at the sum or difference frequency is given by

component at 
$$|\omega_1 \pm \omega_2| = A_1 A_2 \frac{\mu_0 C'_{ox} W}{2L}$$
 (7.35)

and it is seen to be independent of bias conditions. This principle is used for example in [King 97, Borre 97].

As an example, assume that the single-transistor mixer is used as an upconverter. The transistor is processed in the  $0.7\mu m$  process. The size of the transistor is  $W=320\mu m$  and  $L=14\mu m$ . The transistor is biased in the strong inversion region and  $V_{DS}=0V$ . A local oscillator signal is applied between the gate and the source with an amplitude of 0.316V, which corresponds to 0dBm. The baseband signal is a sinusoidal differential signal with an amplitude of 1V. It is applied between the drain and source terminals of the transistor. The output of interest is the component of the drain current at the frequency that is the sum of the baseband frequency and the local oscillator signal. With these data it is found that the amplitude of this component is 0.344mA.

The mixer operation can be interpreted as follows: in strong inversion, a conductive channel exists between drain and source. When an AC voltage  $v_{ds}$  is applied over the channel, an AC current from drain to source results. When the gate bias varies according to a second AC voltage, then the conductivity of this channel is modulated and the AC drain current is modulated by this second AC voltage as well.

In Chapter 8 this mixer type will be studied more in detail.

#### 7.4.1.2 Formulation of the current in terms of $v_{GB}$ , $v_{DB}$ and $v_{SB}$

Equation (7.25) can be rewritten in terms of the voltages  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$  as follows:

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left( \left( v_{GB} - V_{FB} - \phi - \gamma \sqrt{\phi + v_{SB}} \right) \left( v_{DB} - v_{SB} \right) - \frac{1}{2} \left( v_{DB}^2 - v_{SB}^2 \right) \right)$$
(7.36)

It is seen that this equation is not symmetric with respect to  $v_{DB}$  and  $v_{SB}$ . This is not correct: if the potentials at the source and drain are interchanged, then the only difference should be that the drain current changes sign [Tsiv 88]. The asymmetry in equation (7.36) will lead to errors in the derivatives. The reason for this asymmetry is that the threshold voltage has been considered as being constant over the whole channel, and it has been taken equal to the threshold voltage at the source. This assumption corresponds to a uniform depletion region below the inversion layer. Nevertheless, this assumption is widely used. In the next section we will consider a drain current model that takes into account the variation of the depletion layer along the channel. This model is symmetric with respect to source and drain.

## 7.4.2 Nonuniform depletion layer

If the threshold voltage is not considered as a constant over the channel, then the integration of equation (7.19) with a depletion layer charge that is variable along the channel (equation (7.24)) yields the current in terms of  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ :

$$i_{D} = \frac{\mu_{0}C'_{ox}W}{L} \left\{ \left( v_{GB} - V_{FB} - \phi \right) \left( v_{DB} - v_{SB} \right) - \frac{1}{2} \left( v_{DB}^{2} - v_{SB}^{2} \right) - \frac{2}{3}\gamma \left[ \left( \phi + v_{DB} \right)^{3/2} - \left( \phi + v_{SB} \right)^{3/2} \right] \right\}$$
(7.37)

Equation (7.37) clearly shows the symmetry of source and drain. It can be written in the form

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left[ g \left( v_{GB}, v_{DB} \right) - g \left( v_{GB}, v_{SB} \right) \right]$$
 (7.38)

with the function  $g(v_{GB}, v)$  given by

$$g(v_{GB}, v) = (v_{GB} - V_{FB} - \phi) v - \frac{1}{2}v^2 - \frac{2}{3}\gamma (\phi + v)^{3/2}$$
(7.39)

From the expression current we can draw the following conclusions for the derivatives of the drain current:

- expressions for the derivatives of  $i_D$  with respect to  $v_{DB}$  are opposite to the expressions of the derivatives with respect to  $v_{SB}$ .
- derivatives of  $i_D$  with respect to  $v_{GB}$  of order higher than one are zero since the function g is linear with respect to  $v_{GB}$ .
- derivatives of the form  $\partial^n i_D/\partial v_{DB}^m \partial v_{SB}^{(n-m)}$  are zero.

Moreover, since the function  $g(v_{GB},v)$  is quite simple, the expressions of the derivatives are quite simple as well and the power series coefficients defined in equation (7.4) can be derived easily. Expressions for these coefficients are listed in Table 7.5. This table also contains numerical values of the nonlinearity coefficients evaluated for an n-MOS transistor with an effective channel length of  $14\mu m$  and a width of  $320\mu m$ , biased with a gate-bulk voltage of 3.2V, a source-bulk voltage of 1.3V and a drain-source voltage of 0V and 0.45V.

1831			Evaluation at Evaluation at	Evaluation at
- 6	expression in	full	$V_{GB} = 3.2V$	$V_{GB} = 3.2V$
coefficient	terms of $g$	expression	$V_{DB} = 1.3V$	$V_{DB} = 1.75V$
			$V_{SB} = 1.3V$	$V_{SB} = 1.3V$
$g_{mg}$	$\frac{\mu_0 C'_{ox} W}{L} \left( \frac{\partial g(v_{GB}, v_{DB})}{\partial v_{GB}} \right) \\ - \frac{\partial g(v_{GB}, v_{SB})}{\partial v_{GB}} \right)$	$\frac{\mu_0 C'_{ox} W}{L} \left(V_{DB} - V_{SB}\right)$	0~A/V	0.982~mA/V
$K_{2gmg}$	$\frac{\mu_0 C_{ox}' W}{2L} \begin{pmatrix} \partial^2 g(v_{GB}, v_{DB}) \\ \partial v_{GB}^2 \\ -\partial^2 g(v_{GB}, v_{SB}) \end{pmatrix}$	0	$0~A/V^2$	$0~A/V^2$
$K_{3gmg}$	$\frac{\mu_0 C'_{ox} W}{6L} \left( \frac{\partial^3 g(v_{GB}, v_{DB})}{\partial v_{GB}^3} \right) \\ - \frac{\partial^3 g(v_{GB}, v_{SB})}{\partial v_{3.5}^3} \right)$	0	$0~A/V^3$	$0~A/V^3$
$g_{md}$	DB)	$\frac{\mu_0 C_{ox}' W}{L} \left( V_{GB} - V_{FB} - \phi - V_{DB} - \gamma \sqrt{\phi + V_{DB}} \right)$	1.60~mA/V	0.377~mA/V
$K_{2gmd}$	$\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^2 g(v_{GB}, v_{DB})}{\partial v_{DR}^2}$	$-rac{\mu_0 C'_{ox} W}{2L} \left(1 + rac{\gamma}{2\sqrt{\phi + V_{DB}}} ight)$	$-1.37\ mA/V^2$	$-1.37 \ mA/V^2 \ -1.35 \ mA/V^2$
$K_{3gmd}$	$\frac{\mu_0 C'_{ox} W}{6L} \frac{\partial^3 g(v_{GB}, v_{DB})}{\partial v_{DB}^3}$	$\frac{\mu_0 C'_{ox} W}{24L} \frac{\gamma}{(\phi + V_{DB})^{3/2}}$	$22.4~\mu A/V^3$	$16.8~\mu A/V^3$
			continue	continued on next page

Table 7.5: Nonlinearity coefficients of an n-MOS transistor in strong inversion (triode region) in terms of voltages referred to the bulk. They have been computed with the symmetric model of equation(7.37). The function g is defined in equation (7.39). The numerical values have been obtained for a transistor with  $W=320\mu m$  and  $L=14\mu m$  with the model parameters of the  $0.7\mu m$ process (see Table 7.1).

continued from	continued from previous page			
	- AMOUNT MAY		Evaluation at	Evaluation at
	expression in	full	$V_{GB} = 3.2V$	$V_{GB} = 3.2V$
coemcient	terms of $g$	expression	$V_{DB} = 1.3V$	$V_{DB} = 1.75V$
			$V_{SB} = 1.3V$	$V_{SB} = 1.3V$
$g_{ms}$	$\frac{\mu_0 C'_{ox} W}{L} \frac{\partial g(v_{GB}, v_{SB})}{\partial v_{SB}}$	$\frac{\mu_0 C'_{ox} W}{L} \left( V_{GB} - V_{FB} - \phi - V_{SB} - \gamma \sqrt{\phi + V_{SB}} \right)$	1.60~mA/V	1.60~mA/V
$K_{2g_{ms}}$	$\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^2 g(v_{GB}, v_{SB})}{\partial v_{SB}^2}$	$-\frac{\mu_0 C'_{ox} W}{2L} \left( 1 + \frac{\gamma}{2\sqrt{\phi + V_{SB}}} \right)$	$-1.37 \ mA/V^2 \left  -1.37 \ mA/V^2 \right $	$-1.37 \ mA/V^2$
$K_{3g_{ms}}$	$\frac{\mu_0 C'_{ox} W}{6L} \frac{\partial^3 g(v_{GB}, v_{SB})}{\partial v_{SB}^3}$	$rac{\mu_0 C'_{ox} W}{24L} rac{\gamma}{\left(\phi + V_{SB} ight)^{3/2}}$	$22.4 \ \mu A/V^3$	$22.4 \ \mu A/V^3$
$K_{2g_{mg}\&g_{ms}}$	$-\frac{\mu_0 C'_{ox} W}{L} \frac{\partial^2 g(v_{GB}, v_{SB})}{\partial v_{GB} \partial v_{SB}}$	$-rac{\mu_0 C'_{ox} W}{L}$	$-2.18 \ mA/V^2$	$-2.18 \ mA/V^2$
$K_{32g_{mg}\&g_{ms}}$	$-\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^3 g(v_{GB}, v_{SB})}{\partial v_{GB}^2 \partial v_{SB}}$	0	$0 A/V^3$	$0 A/V^3$
$K_{3g_{mg}\&2g_{ms}}$	$-\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^3 g(v_{GB}, v_{SB})}{\partial v_{GB} \partial v_{SB}^2}$	0	$0 A/V^3$	$0 A/V^3$
$K_{2g_{mg}\&g_{md}}$	$\frac{\mu_0 C'_{ox} W}{L} \frac{\partial^2 g(v_{GB}, v_{DB})}{\partial v_{GB} \partial v_{DB}}$	$rac{\mu_0 C'_{ox} W}{L}$	$2.18 \ mA/V^2$	$2.18 \ mA/V^2$
$K_{32g_{mg}\&g_{md}}$	$\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^3 g(v_{GB}, v_{DB})}{\partial v_{GB}^2 \partial v_{DB}}$	0	$0 A/V^3$	$0 A/V^3$
$K_{3g_{mg}\&2g_{md}}$	$\frac{\mu_0 C'_{ox} W}{2L} \frac{\partial^3 g(v_{GB}, v_{DB})}{\partial v_{GB} \partial v_{DB}^2}$	0	$0 A/V^3$	$0 A/V^3$
$K_{2g_{ms}\&g_{md}}$	0	0	$0 A/V^2$	$0 A/V^2$
$K_{32g_{ms}\&g_{md}}$	0	0	$0 A/V^3$	$0 A/V^3$
$K_{3g_{ms}\&2g_{md}}$	0	0	$0 A/V^3$	$0 A/V^3$
$K_{g_{mg}\&g_{ms}\&g_{md}}$	0	0	$0 A/V^3$	$0 A/V^3$

For  $V_{DS}=0V$  the numerical values of the derivatives with respect to  $v_{DB}$  are opposite to the derivatives with respect to  $v_{SB}$ . Indeed, at  $v_{DS}=0V$  the role of source and drain is equivalent. This should be reflected as well in the values of the nonlinearity coefficients. However, since  $g_{ms}$ ,  $K_{2g_{ms}}$  and  $K_{3g_{ms}}$  have been defined as the opposite of the first three derivatives with respect to  $v_{SB}$ , they have the same value (and not the opposite value) as  $g_{md}$ ,  $K_{2g_{md}}$  and  $K_{3g_{md}}$ , respectively. For  $V_{DS} \neq 0V$ , of course, the numerical values are different. In this case, the channel is no longer symmetric with respect to source and bulk.

Most designers are still more familiar with terminal voltages that are referred to the source, instead of the bulk. Equation (7.37) can be rewritten in terms of  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ . This yields

$$i_{D} = \frac{\mu_{0}C'_{ox}W}{L} \left\{ (v_{GS} - V_{FB} - \phi) v_{DS} - \frac{1}{2}v_{DS}^{2} - \frac{2}{3}\gamma \left[ (\phi + v_{SB} + v_{DS})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(7.40)

This equation is implemented in the *level 2 model of SPICE* and it is therefore referred to as the level 2 model equation for the triode region. In addition, the complete level 2 model [Hspi 96] incorporates some small-geometry effects as well, but these effects are discussed later. The nonlinearity coefficients that arise from equation (7.40) are evaluated in the Section 7.4.4, where they are compared to nonlinearity coefficients derived from the simple drain current model of equation (7.25) and from a model in which the variation of the thickness of the depletion layer along the channel is approximated. This model is discussed in the next section.

# 7.4.3 Simplification: linearly varying depletion layer

For numerical simulations of large circuits containing many transistors, the time a computer needs to evaluate the model equations is very critical. Equation (7.40) contains  $\frac{3}{2}$  powers. The evaluation of these powers is expensive from a computational point of view. In several models such as the level 3 model [Liu 82] and the BSIM model [Sheu 87, BSIM 95], the  $\frac{3}{2}$  power dependence has been replaced by a numerical approximation, in order to speed up numerical circuit simulations. For both the level 3 model and the BSIM model the expression of the current (without taking into account small-geometry effects and other effects such as a nonuniform doping) has the form

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left( v_{GS} - V_T - \frac{a}{2} v_{DS} \right) v_{DS}$$
 (7.41)

in which  $V_T$  is given by equation (7.27). The parameter a is found by expanding the  $\frac{3}{2}$  term in equation (7.40) that contains  $v_{DS}$ , into a power series, and keeping the terms in  $v_{DS}$  and  $v_{DS}^2$ . This yields<sup>2</sup>

$$a = \frac{2\sqrt{\phi + v_{SB}} + \gamma}{2\sqrt{\phi + v_{SB}}} \tag{7.42}$$

<sup>&</sup>lt;sup>2</sup>In the Berkeley SPICE implementation an error is made in this series expansion. This error is not corrected in other implementations [Hspi 96, Eldo 91] in order to remain compatible with the SPICE versions of Berkeley.

and for the current

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left[ (v_{GS} - V_{FB} - \phi) v_{DS} - \frac{v_{DS}^2}{2} - \frac{\gamma v_{DS}^2}{4\sqrt{\phi + v_{SB}}} - \gamma \sqrt{\phi + v_{SB}} \cdot v_{DS} \right]$$
(7.43)

As an example, consider the data of the  $0.7\mu m$  and the  $0.5\mu m$  process of Table 7.1. For  $v_{SB}=1.3V$  we find that the value of the parameter a of equation (7.42) is about the same for the two processes, and it equals 1.25.

In the rest of this chapter, equation (7.41) or (7.43) will be referred to as the *level 3 model of* **SPICE** although the actual level 3 model takes into account more effects as the ones discussed in this section. For example, the parameter a takes into account small-geometry effects as well. This has also been done in the BSIM model [Sheu 87, BSIM 95]. In addition, a better accuracy of a for long-channel devices is obtained for high values of  $v_{SB}$  by providing an extra correction factor<sup>3</sup>. Finally, in hand calculations the model derived with a uniform depletion layer is extended with a parameter  $\delta$  as follows [Tsiv 88]:

$$i_D = \frac{\mu_0 C'_{ox} W}{L} \left[ (v_{GS} - V_T) v_{DS} - \frac{1}{2} (1 + \delta) v_{DS}^2 \right]$$
 (7.44)

If  $\delta$  is set to zero, then we obtain equation (7.25). When  $\delta$  is set to  $\gamma/\left(2\sqrt{\phi+v_{SB}}\right)$  then we obtain equation (7.41). For hand calculations, however,  $\delta$  is often treated as a constant. A good compromise between accuracy and simplicity for not too high values of  $v_{DS}$  and  $v_{SB}$  is to set  $\delta$  equal to  $\gamma/(4\sqrt{\phi})$  [Tsiv 88].

The Taylor series approximation made in the level 3 model and the BSIM model or in equation (7.44) corresponds to the simplification that the contribution of the depletion layer charge to the value of the inversion layer charge varies linearly along the channel. Indeed, if expression (7.24) for the depletion layer charge along the channel is expanded into a Taylor series around  $v_{SB}$  and only the linear term is kept, then one finds, using the parameter a

$$Q_B' \approx -C_{ox}' \left( \gamma \sqrt{\phi + v_{SB}} + (a - 1) \left( v_{CB} - v_{SB} \right) \right) \tag{7.45}$$

and the corresponding charge in the inversion layer

$$Q_{I}' = -C_{ox}' \left( v_{GB} - v_{SB} - V_{FB} - \phi - \gamma \sqrt{\phi + v_{SB}} - a \left( v_{CB} - v_{SB} \right) \right)$$
(7.46)

If this value is used to compute the drain current with the integral of equation (7.19), then one will obtain expression (7.41).

In fact, the linear approximation is a compromise between a uniform depletion layer along the channel on one hand, and a more precise variation on the other hand, given in equation (7.24).

The approximation made above is one of the many examples where accuracy is sacrificed for computational speed. The error with this approximation is small for the current. An evaluation of the error for the derivatives will be given in the next section. However, one can already see in

<sup>&</sup>lt;sup>3</sup>This correction factor had a fixed value in the early versions of the BSIM model but in version 3 this can be fitted using the parameter *KETA*.

advance that with the approximation given in equation (7.43) or equation (7.44) the third-order nonlinearity coefficient  $K_{3g_o}$  is zero, since the dependence of the current on  $v_{DS}$  is quadratic. This is not correct as can be seen from the more accurate expression (equation (7.40)) which contains a  $\frac{3}{2}$  power term that comprises  $v_{DS}$ .

## 7.4.4 Comparison of nonlinearity coefficients

Expressions for the nonlinearity coefficients that are derived from the accurate level 2 expression for the drain current, equation (7.37) or equation (7.40), are quite complicated. Nevertheless, some interesting common points and differences can be pointed out with the equations derived with a uniform or a linearly varying depletion layer. First, it is seen that the dependence of the drain current on  $v_{GS}$  is identical for the levels 1, 2 and 3. Hence, all nonlinearity coefficients that depend upon a derivative with respect to  $v_{GS}$  including cross-derivatives that contain at least one differentiation with respect to  $v_{GS}$  are the same for the three models.

Table 7.6 lists the values of the nonlinearity coefficients as they are derived from the accurate level 2 expression of equation (7.37). The values are compared to the values obtained with the simple level 1 model of equation (7.25) and with the level 3 model (equation (7.43)). The nonlinearity coefficients for the level 1 model can be found using the expressions listed in Table 7.4.

The nonlinearity coefficients that comprise a derivative with respect to  $v_{GS}$  are not listed, since they are the same for the different models, as mentioned above. The nonlinearity coefficients have been evaluated for the same transistor dimensions and operating points as used in Tables 7.4 and 7.5.

The error between the nonlinearity coefficients obtained with the level 2 model and with the other two models is computed as follows. When the nonlinearity coefficients obtained with the three models are all found to be zero, then the error is 0%. Else, when a nonlinearity coefficient computed with the level 2 model is not zero, then the signed relative error is given by

signed relative error [%] = 
$$\frac{(value found with level 1 or 3) - (value found with level 2)}{(value found with level 2)} \cdot 100$$
(7.47)

As a result the error can be negative. When a nonlinearity coefficient obtained with the level 2 model is found to be nonzero, while it is found to be zero with one of the other two models, then the error is -100%.

It is seen that the level 3 model is a more accurate approximation than the level 1 model. Indeed, for the level 3 model the error on the nonlinearity coefficients is quite small except for the coefficients  $K_{3g_{mb}\&2g_o}$  and  $K_{3g_o}$ . For the latter coefficient both level 1 and level 3 yield a value of zero. This is not correct: the value differs from zero due to the  $\frac{3}{2}$  power dependence of the drain current on  $v_{DS}$  (see equation (7.37)). Also it is seen that the error on the nonlinearity coefficients obtained with the level 1 and level 3 models, increase with the order of the derivative. For example, the relative error on the derivatives of the current with respect to  $v_{SB}$  increases from 4.9% to 27% for level 1 and from 0.83% to 6.9% for level 3.

	evaluated at	,		evaluated at		
	$V_{GS} = 1.9V$	error	error	$V_{GS} = 1.9V$	error	error
coeff.	$V_{DS} = 0V$	with	with	$V_{DS} = 0.45V$	with	with
COCH.	$V_{SB} = 1.3V$	"level 1"	"level 3"	$V_{SB} = 1.3V$	"level 1"	"level 3"
	"level 2"	model	model	"level 2"	model	model
	model			model		
$i_D$	0A	0%	0%	0.444mA	12%	0.45%
$g_{mb}$	0A/V	0%	0%	0.242mA/V	4.9%	-0.83%
$K_{2g_{mb}}$	$0A/V^2$	0%	0%	$-26.1 \mu A/V^2$	-16%	-2.7%
$K_{3g_{mb}}$	$0A/V^3$	0%	0%	$5.67 \mu A/V^3$	27%	-6.9%
$ g_o $	1.60mA/V	0%	0%	0.377mA/V	64%	3.4%
$K_{2g_o}$	$-1.37mA/V^2$	-20%	0%	$-1.35mA/V^{2}$	19%	1.5%
$ K_{3q_{\alpha}} $	$22.4\mu A/V^{3}$	-100%	-100%	$16.8 \mu A/V^3$	-100%	-100%
$K_{2_{q_m},\&q_q}$	$-0.565mA/V^2$	0%	0%	$-0.513mA/V^2$	10%	1.7%
$R_{32g_{mb}\&g_{o}}$	$67.3\mu A/V^{3}$	0%	0%	$50.3\mu A/V^{3}$	34%	9.4%
$K_{3_{g_{mb}\&2g_o}}$	$67.3\mu A/V^3$	-100%	0%	$50.3\mu A/V^3$	-100%	34%

Table 7.6: Nonlinearity coefficients of the drain current in the triode region, obtained with the accurate model of level 2 (equation (7.37)). The relative errors on the nonlinearity coefficients obtained with the models of level 1 (equation (7.25)) and with the level 3 model (equation 7.43) are given as well. Transistor dimensions are  $W=320\mu m$  and  $L=14\mu m$ . The model parameters are the ones from the  $0.7\mu m$  process listed in Table 7.1. Nonlinearity coefficients that comprise a derivative with respect to  $v_{GS}$  are not listed since they are the same as in Table 7.4.

# 7.5 Drain current in saturation without small-geometry effects

When the drain-source voltage  $v_{DS}$  of a transistor in the triode region increases, the drain end of the channel becomes less inverted. From a certain value of  $v_{DS}$  the drain end is no longer strongly inverted. In other words, the inversion layer gets pinched off. For even larger values of  $v_{DS}$ , a larger part of the channel is pinched off. Then the transistor is said to be in the saturation region. This situation is shown in Figure 7.5. The value of  $v_{DS}$  at which the transition from the triode region to the saturation region occurs, is denoted as the saturation voltage  $v_{DSAT}$ .

The above representation is very simplified and it leads to physical inconsistencies, as we shall see below. Nevertheless, it gives an initial idea about the transistor operation in saturation.

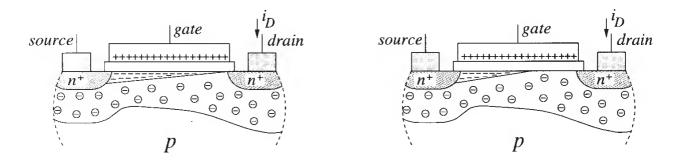


Figure 7.5: Simplified schematic representation of the situation at the onset of saturation (left) where  $v_{DS} = v_{DSAT}$  and for  $v_{DS} > v_{DSAT}$  (right).

An exact computation of the current close to the onset of saturation is difficult. The reason is that the channel at the drain end is no longer strongly inverted, and the basic equations of Section 7.3 that rely on strong inversion operation along the complete channel are no longer valid. Nevertheless, it is common practice to use the strong inversion relationships as a starting point for the derivation of  $v_{DSAT}$ . The resulting error on  $v_{DSAT}$  is small in many applications [Tsiv 88].

A common practice to obtain  $v_{DSAT}$  is to find the value of  $v_{DS}$  for which the drain current becomes independent of  $v_{DS}$ . In other words,  $v_{DSAT}$  is found by solving the equation

$$\frac{\partial i_D}{\partial v_{DS}} = 0 ag{7.48}$$

for  $v_{DS}$ . This corresponds to the situation that at the onset of saturation, the charge  $Q_I'$  of the inversion layer at the drain end becomes zero. In fact,  $Q_I'=0$  is not correct, since the carriers would then have to travel with infinite drift velocity through the pinched-off region in order to obtain a nonzero current.

Just as for the triode region, the nonlinearity coefficients will first be derived for the simplifying assumption that the depletion layer underneath the channel is uniform along the channel, after which the variation of the thickness of this depletion layer is taken into account.

#### 7.5.1 Uniform depletion layer

When equation (7.25) is used for the expression of the drain current, then setting  $\partial i_D/\partial v_{DS}$  to zero and solving for  $v_{DS}$  yields the following value for  $v_{DSAT}$ :

$$v_{DSAT} = v_{GS} - V_T \tag{7.49}$$

The current in the saturation region is obtained by substituting  $v_{DSAT}$  for  $v_{DS}$  in the current equation for the triode region (equation (7.25)). Doing so, one obtains

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{L} \left[ (v_{GS} - V_T) v_{DSAT} - \frac{v_{DSAT}^2}{2} \right]$$
 (7.50)

$$=\frac{\mu_0 C'_{ox} W}{2L} \left(v_{GS} - V_T\right)^2 \tag{7.51}$$

This expression is seen to be independent of  $v_{DS}$ . However, this is not correct. When  $v_{DS}$  increases, a larger portion of the channel gets pinched off. As a result, the length of the inverted part of the channel decreases. Since the current is inversely proportional to the effective channel length, it increases as  $v_{DS}$  increases. This effect, denoted as **channel length modulation** is very often modeled by multiplying the current of equation (7.51) with a factor  $(1 + \lambda v_{DS})$  of  $1/(1 - \lambda v_{DS})$ , in which  $\lambda$  is called the **channel-length modulation factor** [Anto 88, Lak 94] Hspi 96]. Doing so, the drain current becomes

$$i_D = i_{DSAT} (1 + \lambda v_{DS}) = \frac{\mu C'_{ox} W}{2L} (v_{GS} - V_T)^2 (1 + \lambda v_{DS})$$
 (7.52)

The current  $i_{DSAT}$  is the value that is found when the dependence of the current on  $v_{DS}$  is neglected. With this model for channel-length modulation the drain current is linearly dependent on  $v_{DS}$ . As a result, the first derivative of the drain current with respect to  $v_{DS}$  is a constant. This derivative, which is nothing else but the output conductance is given by

$$g_o = \frac{\partial i_D}{v_{DS}} = \lambda \cdot i_{DSAT} \tag{7.53}$$

A constant  $g_o$  means that the slope of the drain current in the transistor characteristic is assumed to be constant, as shown in Figure 7.6. A modeling of the channel-length modulation in this way is very poor [Tsiv 88]. In the level 2 model of SPICE channel-length modulation is modeled with  $\lambda$  as well, but instead of multiplying the current expression without channel-length modulation with a factor  $(1 + \lambda . v_{DS})$ , a division by  $(1 - \lambda v_{DS})$  is performed. By this division the slope is no longer constant even if  $\lambda$  is a constant. Nevertheless, this model is not accurate enough for a good modeling of the output conductance [Hspi 96, Tsiv 93b]<sup>4</sup>. Many other models also give rise to a poor model for  $g_o$  in the saturation regime, such as the level 3 model [Tsiv 93b, Hspi 96] and the early BSIM models [Sheu 87], as illustrated in [Gow 91, Hspi 96]. The discussion of a

<sup>&</sup>lt;sup>4</sup>In the level 2 model it is possible, depending on the specified parameters, to let  $\lambda$  either be a constant, or a more complicated expression, but in both cases a poor model for the output conductance results.

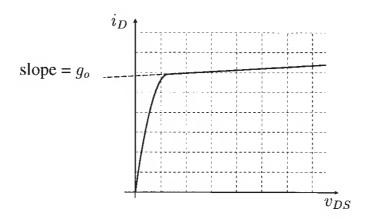


Figure 7.6: Simplified transistor characteristic with a constant slope in the saturation regime.

more accurate modeling of the output conductance is postponed until Section 7.11. Meanwhile, the simple model of the output conductance according to equation (7.53) will be used.

Equation (7.52) is the expression for the drain current in saturation as it is implemented in the SPICE level 1 model [Anto 88, Hspi 96]. Although the accuracy is poor for modern MOS transistors, this equation is still often used for hand calculations thanks to its simplicity. Equation (7.52) reveals that the drain current depends on the square of  $v_{GS}$ . This is why this model equation is often referred to as the *square-law model* for MOS transistors.

The nonlinearity coefficients that are obtained with the quadratic model have already been given in Table 3.2.

In Table 3.2 it is seen that, due to the quadratic dependence of the drain current on  $v_{GS}$  the third-order derivative of the drain current with respect to  $v_{GS}$  is zero. Hence

$$K_{3q_m} = 0 (7.54)$$

This is an oversimplification as we shall see in later sections. The second-order nonlinearity coefficient  $K_{2g_m}$  is not zero. The normalized second-order nonlinearity coefficient  $K'_{2g_m}$  is given by

$$K'_{2g_m} = \frac{1}{2\left(V_{GS} - V_T\right)} \tag{7.55}$$

For  $V_{GS} - V_T = 0.2V$  this coefficient equals  $2.5V^{-1}$ .

It is interesting to note that this value is much smaller than the value of the normalized nonlinearity coefficient  $K_{2g_m}'$  for a bipolar transistor. In Section 6.2 this was found to be about  $1/(2V_t)$  under low injection conditions, or about  $20V^{-1}$  at room temperature. In contrast with a bipolar transistor the normalized nonlinearity coefficient  $K_{2g_m}'$  of a MOS transistor can still be adapted by the bias conditions: it decreases as  $V_{GS} - V_T$  increases.

The simple modeling of the channel-length modulation using  $\lambda$  involves a  $g_o$  that is constant with respect to  $v_{DS}$ , and, hence,  $K_{2g_o}$  and  $K_{3g_o}$  are zero. This is not realistic, as will be explained in Section 7.11.

Consider now  $g_{mb}$  and its derivatives. The normalized nonlinearity coefficients  $K'_{2g_{mb}}$  and  $K'_{3g_{mb}}$  are easily derived from Table 3.2. They are given

$$K'_{2g_{mb}} = -\frac{1}{4} \frac{1}{(\phi + V_{SB})} \left( 1 + \frac{\gamma \sqrt{\phi + V_{SB}}}{V_{GS} - V_T} \right)$$
 (7.56)

$$K_{3g_{mb}}' = \frac{1}{8} \frac{1}{(\phi + V_{SB})^2} \left( 1 + \frac{\gamma \sqrt{\phi + V_{SB}}}{V_{GS} - V_T} \right)$$
 (7.57)

Compared to the corresponding normalized nonlinearity coefficients in the triode region (equations (7.30) and (7.31)), it is seen that the absolute value of the coefficients is somewhat larger here.

## 7.5.2 Nonuniform depletion layer

If variation of the depletion layer along the channel is taken into account, then the current in the saturation regime needs to be derived from equation (7.37). Following the same procedure as above, one finds for  $v_{DSAT}$  [Tsiv 88]

$$v_{DSAT} = v_{GS} - \phi - V_{FB} + \frac{\gamma^2}{2} - \gamma \sqrt{v_{GS} - V_{FB} + v_{SB} + \frac{\gamma^2}{4}}$$
 (7.58)

Substitution of this value into equation (7.40) yields the expression for the current in the saturation regime

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{L} \left\{ \frac{(v_{GS} - V_{FB} - \phi)^2}{2} - h(v_{GS}, v_{SB}) \right\} (1 + \lambda v_{DS})$$
(7.59)

in which  $h(v_{GS}, v_{SB})$  is given by

$$h(v_{GS}, v_{SB}) = \frac{\gamma^4}{4} - \frac{\gamma^3}{2} \sqrt{v_{GS} - V_{FB} + v_{SB} + \frac{\gamma^2}{4}} + \frac{\gamma^2}{2} (v_{GS} - V_{FB} + v_{SB}) + \frac{2}{3} \gamma \left[ (\phi + v_{SB} + v_{DSAT})^{3/2} - (\phi + v_{SB})^{3/2} \right]$$
(7.60)

Equation (7.59) is implemented in the level 2 model of SPICE.

It is seen that  $h(v_{GS}, v_{SB})$  is zero if  $\gamma$  is zero. Hence, if the body-effect coefficient is very small, then the transistor operates as a square-law device. For scaled devices, however, this is no longer the case, since the body-effect coefficient becomes relatively more important [Tsiv 88].

From equation (7.59) it is seen that the dependence of the drain current on  $v_{GS}$  is not purely quadratic anymore. This is due to the second term in  $h(v_{GS}, v_{SB})$  which contains a square root that comprises  $v_{GS}$  and to the  $\frac{3}{2}$  power that contains  $v_{DSAT}$ . Also,  $v_{DSAT}$  is no longer linearly dependent on  $v_{GS}$  as can be seen from equation (7.58). As a result, the third-order derivative of the drain current with respect to  $v_{GS}$  is no longer zero such that  $K_{3g_m}$  is not zero anymore and third-order distortion will arise.

## 7.5.3 Simplification: linearly varying depletion layer

In this section the drain current in the saturation region is derived for the simplifying assumption that the thickness of the depletion layer changes linearly along the channel. This assumption is made in the level 3 model. Also, the approximation made in the BSIM model is discussed briefly as well.

In order to obtain the level 3 equation for the drain current in saturation, one must first obtain the value for  $v_{DSAT}$ . Following the same approach as above, one must start from the expression for the current in the triode region (equation (7.41)), compute  $\partial i_D/\partial v_{DS}$ , set this to zero and solve for  $v_{DS}$ . This yields

$$v_{DSAT} = \frac{v_{GS} - V_T}{a} \tag{7.61}$$

with a given in equation (7.42). The parameter a only depends on  $v_{SB}$  and hence  $v_{DSAT}$  is linearly proportional to  $v_{GS}$ . The value for  $v_{DSAT}$  obtained in this way must then be substituted in the drain current expression for the triode region, which yields

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{2aL} \left( v_{GS} - V_T \right)^2 \tag{7.62}$$

It is seen that the current in this model again depends on the square of  $v_{GS} - V_T$ , just as with the simplified model that relies on a uniform depletion layer along the channel (equation (7.51)).

The expression for the current in saturation with the BSIM model is derived by stating that at the onset of saturation the velocity of carriers at the drain end becomes equal to the saturation velocity. This will be discussed further in Section 7.7. For long-channel transistors, the value of  $v_{DSAT}$  is also given by equation (7.61), with, however, an adapted value of a in order to have a better accuracy for high values of  $v_{SB}$  (see also Section 7.4.3). Consequently, the current in the saturation region for large transistors (and neglecting channel-length modulation) also depends on the square of  $v_{GS} - V_T$ .

# 7.5.4 Comparison of nonlinearity coefficients

Table 7.7 lists the numerical values of the nonlinearity coefficients that have been derived using the accurate model of equation (7.59), denoted in this chapter as the level 2 model. The values have been obtained for an n-MOS transistor with  $W=320\mu m$  and  $L=14\mu m$ , fabricated in the  $0.7\mu m$  process with the model parameters of Table 7.1. Also, the deviations with the values obtained with the simple quadratic model (level 1) and with the level 3 model are listed. The modeling of channel-length modulation has been omitted. As a result, no relevant values are obtained for nonlinearity coefficients that comprise derivatives with respect to  $v_{DS}$ . The discussion on modeling the output conductance in saturation is postponed until Section 7.11.

It should be noted that the values for the level 1 model can be found by using the expressions listed in Table 3.2, in which  $\lambda$  is set to zero, in order to neglect the effect of channel-length modulation.

coefficient	value ("level 2" model)		with	viation "level 1"	deviation with "level 3" model		
$i_{DSAT}$	0.470	mA	31	%	-0.85	%	
$g_m$	1.29	mA/V	30	%	-1.5	%	
$K_{2g_m}$	0.888	$mA/V^2$	29	%	-2.5	%	
$K_{3g_m}$	8.34	$\mu A/V^3$	-100	%	-100	%	
$g_{mb}$	0.313	mA/V	39	%	-2.2	%	
$K_{2g_{mb}}$	-79.3	$\mu A/V^2$	63	%	6.6	%	
$K_{3g_{mb}}$	14.1	$\mu A/V^3$	118	%	-23	%	
$K_{2_{g_m}\&g_{mb}}$	-0.406	$mA/V^2$	46	%	-4.9	%	
$K_{3_{2g_m}\&g_{mb}}$	25.0	$\mu A/V^3$	-100	%	70	%	
$K_{3_{g_m}\&2g_{mb}}$	25.0	$\mu A/V^3$	183	%	-51	%	

Table 7.7: Nonlinearity coefficients for the drain current in the saturation regime, according to the accurate model of equation (7.59). Values are computed for an n-MOS transistor with  $W=320\mu m,\ L=14\mu m,\ V_{GS}=1.9V,\ V_{DS}=1.45V,\ V_{SB}=1.3V$  and with the model parameters of the  $0.7\mu m$  (See Table 7.1). Also, the deviations of the level 1 model and the level 2 model are listed.

From Table 7.7 it is seen that the nonlinearity coefficients obtained with the level 1 mode deviate considerably from the values obtained with the level 2 model. The errors for the level 3 model are usually smaller. However, it is seen that with the level 3 model  $K_{3g_m}$  is zero. This is due to the simplification in the level 3 equations, that give rise to a quadratic dependence of the current on  $v_{GS} - V_T$ . On the other hand, the lower-order derivatives of the drain current with respect to  $v_{GS}$ , corresponding to  $g_m$  and  $K_{2g_m}$ , are much more accurate than with the level 1 model.

Other nonlinearity coefficients for which the "level 3" approximation yields large errors are  $K_{3_{2g_m\&g_{mb}}}$  and  $K_{3_{g_m\&2g_{mb}}}$ .

# 7.6 Effective mobility

Until now we have set the surface mobility  $\mu$  equal to  $\mu_0$ , which is constant along the channel and independent of the applied voltages. In reality, the mobility of electrons is reduced by several effects and it depends on the position in the channel and on the applied voltages.

The electric field that is perpendicular to the direction of the nominal current flow, referred to as the normal field or vertical field, accelerates the electrons towards the interface between silicon and the gate oxide, where they suffer collisions in addition to the collisions with the crystal lattice and with ionized impurity atoms. These collisions reduce the mobility of the electrons. This effect is more severe for modern technologies than for older ones. The reason is that the thickness of the gate oxide scales down as the minimum channel length decreases [Tsiv 88]. Due to this thinner oxide the normal or vertical electric field is higher. In this way, mobility reduction due to the normal field is often considered as a small-geometry effect.

In literature many models have been published for the mobility of electrons in the inversion layer [Sab 79, Schw 83, Liang 86, Hua 90]. In these models the mobility reduction is modeled in an empirical or semi-empirical way by taking into account different physical phenomena, such as Coulomb scattering, scattering due to the surface roughness and phonon scattering. Each of these effects contributes to the reduction of the mobility.

### 7.6.1 Mobility model of Sabnis and Clemens

In [Sab 79] a relationship is found between the surface mobility and the average normal electric field  $E_{eff}$  that is experienced by carriers in the inversion layer. This relationship only holds if the mobility is limited by phonon scattering. The relationship fails at low temperatures (e.g. 77K) when coulombic scattering becomes important [Hua 90].

The normal field, averaged in the vertical direction, is given by [Sab 79, Tsiv 88, Koh 89]

$$E_{eff} = -\frac{Q_I' + Q_B'}{\varepsilon_{Si}} \tag{7.63}$$

in which  $Q'_I$  and  $Q'_B$  are given by equation (7.20) and equation (7.24), respectively.

Many empirical forms have been formulated to model the universal relationship between mobility and average normal field. A formulation on which many transistor models are based is given by

$$\mu = \frac{\mu_0}{1 + \alpha_\theta E_{eff}} \tag{7.64}$$

in which  $\mu_0$  and  $\alpha_\theta$  are fit parameters [Tsiv 88]. Note that  $\mu_0$  is the same parameter as the one used in the previous sections to model the mobility.

In equation (7.63) it is seen that the average normal field depends on the depletion layer charge and the inversion layer charge. Since these charges depend on the position in the channel, it is clear that the mobility depends on the position in the channel or on  $v_{CB}$ , the voltage difference between a point in the channel and the bulk.

We recall that the surface mobility  $\mu$  appears in the integrand of the expression for the dracurrent, equation (7.19). Since  $\mu$  is no longer seen as a constant, it cannot be moved outside the integral, as was done in the previous sections. One can try to determine  $\mu$  as a function  $v_{CB}$  and then compute the integral in equation (7.19) to obtain an expression for the current. It general, this is very complex and, depending on the model for  $\mu$ , it might lead to an integral the cannot be solved analytically [Groen 94].

An alternative approach which is widely used is to define an *effective mobility*  $\mu_{eff}$  such the when it is used in the expression

$$i_D = \frac{W}{L} \mu_{eff} \int_{v_{SB}}^{v_{DB}} (-Q_I') \, dv_{CB}$$
 (7.6)

 $i_D$  is predicted correctly. Since the electric field and hence  $\mu$  at each point in the channel deper on the terminal voltages, it is expected that to make equation (7.65) give the same result equation (7.19),  $\mu_{eff}$  will be a function of the terminal voltages. In the next section, sever models for this function will be presented that are derived from the mobility formulation equation (7.64).

## 7.6.2 Drain current in the triode region

Before presenting the drain current expression, the effective mobility  $\mu_{eff}$  will be computed. The assumptions and simplifications made in this computation will of course influence the accurace on the drain current expression, and, even more, the accuracy on the derivatives. Initially, we will take into account the variation of the depletion layer along the channel without simplification, a we did in Section 7.4.2.

Using the mobility model of equation (7.64), one can prove that  $\mu_{eff}$  can be approximate by [Merck 72, White 80, Gar 87a, Tsiv 88]

$$\mu_{eff} \approx \frac{\mu_0}{1 + \theta f_\mu} \tag{7.66}$$

where  $f_{\mu}$  is a function of the transistor terminal voltages that will be discussed below, and is an empirical parameter, the so-called *mobility-reduction coefficient*. It can be approximate by [Gar 87a]

$$\theta = \frac{\mu_0}{2t_{or}v_{norm}} \tag{7.67}$$

in which  $v_{norm}$  is a proportionality constant with the dimensions of velocity. It is approximately equal to  $2.2 \times 10^9 \ cm/s$ . Equation (7.67) reveals that  $\theta$  increases as the gate oxide thickness decreases. Since  $t_{ox}$  decreases with scaling,  $\theta$  will increase. In other words, mobility reduction will become more important when devices scale down. Apart from equation (7.67) other empirical equations for  $\theta$  are used [Tsiv 88], but the dependence on  $t_{ox}$  is the same.

For the  $0.7\mu m$  process we find from Table 7.1 that  $\mu_0 = 0.047 m^2/(V.s)$  and  $t_{ox} = 1.7 \times 10^{-8}$  Equation (7.67) then yields  $\theta = 0.062 V^{-1}$ . This does not correspond to the value of  $0.079 V^{-1}$ 

from Table 7.1. The reason is that the latter value is a fit parameter rather than a more physical parameter. However, the same argument could be given for the parameter  $v_{norm}$ , but this parameter has been determined by measurements with different technologies with different oxide thicknesses [Gar 87a].

The function  $f_{\mu}$  in equation (7.66) is a function of the transistor terminal voltages  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$  [White 80, Tsiv 88]. It is found to be

$$f_{\mu} = (v_{GB} - V_{FB} - \phi) - \frac{1}{2}(v_{DB} + v_{SB}) + \frac{2}{3}\gamma \frac{(\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2}}{v_{DB} - v_{SB}}$$
(7.68)

or, in terms of  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ 

$$f_{\mu} = (v_{GS} - V_{FB} - \phi) - \frac{1}{2}v_{DS} + \frac{2}{3}\gamma \frac{(\phi + v_{SB} + v_{DS})^{3/2} - (\phi + v_{SB})^{3/2}}{v_{DS}}$$
(7.69)

With the value of  $\mu_{eff}$  from equation (7.66), an expression for the current can be derived. It is obtained by replacing the mobility  $\mu_0$  with  $\mu_{eff}$  in previous expressions of the drain current in the triode region. Since we assume in this section that the depletion layer charge varies along the channel, as we did in Section 7.4.2, the value for  $\mu_{eff}$  given in equation (7.66) is to be used in conjunction with the accurate model equation for the drain current (equation (7.37) or (7.40)) that takes into account the variation of the depletion layer charge as well. In this way, the current in the triode region in terms of voltages referred to the bulk, is given by

$$i_{D} = \frac{\mu_{0}}{1 + f_{\mu}} C'_{ox} \frac{W}{L} \left\{ (v_{GB} - V_{FB} - \phi) (v_{DB} - v_{SB}) - \frac{1}{2} (v_{DB}^{2} - v_{SB}^{2}) - \frac{2}{3} \gamma \left[ (\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(7.70)

with  $f_{\mu}$  given by equation (7.68). The only difference of this expression with the expression that does not take into account mobility reduction (equation (7.37)) is the factor  $(1 + f_{\mu})$  in the denominator.

When voltages are referred to the source, then the drain current is given by

$$i_{D} = \frac{\mu_{0}}{1 + f_{\mu}} C'_{ox} \frac{W}{L} \left\{ (v_{GS} - V_{FB} - \phi) v_{DS} - \frac{1}{2} v_{DS}^{2} - \frac{2}{3} \gamma \left[ (\phi + v_{SB} + v_{DS})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(7.71)

with  $f_{\mu}$  now given by equation (7.69).

Symmetry of source and drain We recall that the accurate "level 2" expression of the current in the triode region is symmetric with respect to source and drain, in the sense that, when the roles of  $v_{SB}$  and  $v_{DB}$  are interchanged, then the current simply changes sign. On the other hand, it is seen that, when in the expression for  $f_{\mu}$  in terms of voltages referred to the bulk (equation (7.68)), the roles of  $v_{SB}$  and  $v_{DB}$  are interchanged, then  $f_{\mu}$  does not change. As a result, the drain current

expression (7.70) exhibits the same symmetry as the expression that does not take into account mobility reduction.

For later use we rewrite the drain current as follows

$$i_{D} = \frac{\mu_{0}C'_{ox}W}{L} mobred\left(v_{GB}, v_{DB}, v_{SB}\right) \left[g\left(v_{GB}, v_{DB}\right) - g\left(v_{GB}, v_{SB}\right)\right]$$
(7.72)

in which  $mobred(v_{GB}, v_{DB}, v_{SB})$  is given by

$$mobred(v_{GB}, v_{DB}, v_{SB}) = \frac{1}{1 + f_{\mu}}$$
 (7.73)

The function  $g(v_{GB}, v_{SB})$  is given by equation (7.39) and  $f_{\mu}$  is found in equation (7.68).

When the expression of  $f_{\mu}$  is evaluated for  $v_{DS}=0V$  or  $v_{DB}=v_{SB}$ , then the last term contains a division by zero. However, the numerator is zero as well, and the value of the last term is well defined. It can be computed by taking the limit of this term for  $v_{DB}=v_{SB}$ . This yields

$$f_{\mu}\Big|_{v_{DB}=v_{SB}} = (v_{GB} - V_{FB} - \phi) - v_{SB} + \gamma \sqrt{v_{SB} + \phi}$$
 (7.74)

Simplifying assumption The derivation of the expression for the effective mobility has been skipped in this book. Interested readers are referred to [Merck 72, White 80, Tsiv 88]. Nevertheless, it is interesting to mention that in the derivation a simplifying assumption has been made the electric field in the direction of the current flow in the channel is assumed to be constant and given by  $v_{DS}/L$ . This is not exact: the electric field increases along the channel from drain to source when a positive  $v_{DS}$  is applied. The simplifying assumption is only a good approximation for low values of  $v_{DS}$  far from the saturation region. It can be questioned whether this assumption leads to a large error, not only on the current, but also on its derivatives. To this purpose, the current expression (7.70) and its derivatives has been compared to the complicated exact expression [Merck 72] and its derivatives without the simplifying assumption of a constant electric field along the channel. With the model parameters of the  $0.7\mu m$  process of Table 7.1 the error is never larger than 3%.

Further simplification to a uniform depletion layer When the width of the depletion layer is assumed to be uniform along the channel, then  $f_{\mu}$  simplifies to [Merck 72, White 80, Tsiv 88]

$$f_{\mu} = v_{GS} - V_T + 2\gamma \sqrt{\phi + v_{SB}} - \frac{v_{DS}}{2} \tag{7.75}$$

This value still depends on  $v_{DS}$ . In [Pow 92] this term is dropped. Very often,  $f_{\mu}$  is further simplified to

$$f_{\mu} = v_{GS} - V_T \tag{7.76}$$

This is done for example in the level 3 model equation and also in many other models [Merck 72, Klaas 76, White 80, Sodi 84, Gar 87a, Sheu 87, Enz 95]<sup>5</sup>. With this simplification the drain current model for the triode region reduces to

$$i_D = \frac{\mu_0}{1 + \theta \left( v_{GS} - V_T \right)} C'_{ox} \frac{W}{L} \left( \left( v_{GS} - V_T \right) v_{DS} - \frac{v_{DS}^2}{2} \right) \tag{7.77}$$

This equation is very often used for hand calculations. Note that this equation is again not symmetric with respect to source and drain.

It should be noted that the value of  $\theta$  that is used in equation (7.77) is not the same as the value used in equation (7.68): for both equations  $\theta$  is usually determined by fitting measured values for the current in the triode region at low  $v_{DS}$  values and  $v_{SB} = 0V$  onto one of the two equations [Anto 88]. Since the measured data are of course the same for the two equations, while the equations themselves are different, a different value for  $\theta$  will result.

Equation (7.77) contains some oversimplifications. The dependence of the mobility reduction on  $v_{SB}$  is only through  $V_T$ . This means that mobility reduction would become less severe as  $v_{SB}$  increases. This is not correct: an increase of  $v_{SB}$  makes the bulk more negative compared to the source. As a result, the electrons are more pushed into the direction of the surface. This has the same effect as an increase of the gate voltage. Hence,  $\mu_{eff}$  should decrease as  $v_{SB}$  increases. In order to model this effect, a term which is linear in  $v_{SB}$ , is very often added to  $f_{\mu}$  [Chow 92b, Gow 91, Matt 96], which results in an effective mobility

$$\mu_{eff} = \frac{\mu_0}{1 + \theta (v_{GS} - V_T) + \theta_B v_{SB}}$$
 (7.78)

In Section 7.6.5 it will be investigated whether an accurate modeling of the dependence of  $\mu_{eff}$  on  $v_{SB}$  will largely influence the nonlinearity coefficients that correspond to derivatives with respect to  $v_{SB}$ . If so, then the term  $\theta_B v_{SB}$  in equation (7.78) is important.

## 7.6.3 Drain current in the saturation region

For the determination of the current in the saturation regime, a value for  $v_{DSAT}$  needs to be determined first. This is done as before, by putting  $\partial i_D/\partial v_{DS}$  equal to zero and then solving for  $v_{DS}$ . However, if this procedure is applied with the current equation (7.71), then putting  $\partial i_D/\partial v_{DS}$  equal to zero leads to an expression from which  $v_{DSAT}$  cannot be computed analytically. Hence,  $v_{DSAT}$  must be computed by iteration. This approach is not followed in numerical circuit simulators, since iterations require too much computation time. Instead, a simpler expression is used for the current, such that  $v_{DSAT}$  can be computed analytically. Of course, such simplification introduces an error. The error on  $v_{DSAT}$  and the drain current is small, but the error on the nonlinearity coefficients can be high, as we will see in Section 7.6.5.

<sup>&</sup>lt;sup>5</sup>Although the mobility reduction due to the normal field has been derived here starting from the universal mobility model of Sabnis and Clemens, published in 1979 [Sab 79], many transistor models published before 1979 model the mobility reduction in the same way as expression (7.77). The derivations were based on other insights and interpretations.

Although  $v_{DSAT}$  cannot be computed analytically when we start from the complicated drain current model of equation (7.71), it is still possible to compute explicit expressions for the derivatives of the current in terms of the terminal voltages and  $v_{DSAT}$ . This is shown in Appendix F. Expressions obtained in this way will be used to evaluate the nonlinearity coefficients in later sections. The iterations that are required to find  $v_{DSAT}$  require more CPU time than the evaluation of a simplified explicit expression for  $v_{DSAT}$ . This is not a problem: the analog circuits the distortion of which is computed, are fairly small. Moreover, the number of iterations is usually very low if a good initial value is used for  $v_{DSAT}$ . Such value can be obtained from a simpler model that uses an explicit expression for  $v_{DSAT}$ . Also, the enhanced accuracy on the nonlinearity coefficients might be preferred to shorter simulation times with less accurate results for the distortion.

Saturation regime with a simple model of mobility reduction and a uniform depletion layer If the mobility reduction factor  $f_{\mu}$  is independent of  $v_{DS}$ , then the value of  $v_{DSAT}$  is unaffected by mobility reduction. For the simple model of equation (7.25) combined with  $f_{\mu}$  being equal to  $(v_{GS} - V_T)$ , one obtains for the current in saturation (without channel-length modulation)

$$i_{DSAT} = \frac{\mu_0}{1 + \theta(v_{GS} - V_T)} C'_{ox} \frac{W}{L} (v_{GS} - V_T)^2$$
(7.79)

This equation is widely used for hand calculations. With this equation a good agreement is obtained with measurements on the nonlinear relationship between the drain current and  $v_{GS}$  for devices with channel lengths down to  $1\mu m$  [Gar 87b], at least for  $v_{BS}=0V$  and  $v_{GS}$  not too high. The accuracy of this simple model will be further investigated in Section 7.6.5.

# 7.6.4 Other mobility models

Although most modern transistor models implicitly rely on the mobility model of Sabnis and Clemens, as we did in the above sections, some drain current models published in literature use other mobility models. For example, in [Pat 90] the model of Schwarz and Russek [Schw 83] is used in order to compute distortion of a transistor in the triode region. The results for long channels and low gate voltages qualitatively correspond to the results reported in [Groen 94] at low gate voltages. In the same paper, the model of [Hua 90] is used in order to explain distortion measurement results on single-transistor circuits at high gate voltages.

# 7.6.4.1 The mobility model of Frohman-Bentchkowsky

One of the earliest models for the dependence of mobility on the normal field is the one from Frohman-Bentchkowsky that has been implemented in the level 2 model of SPICE. It is given by [Froh 68, Anto 88, Hspi 96]

$$\mu_{eff} = \mu_0 \left[ \frac{UCRIT.\,\varepsilon_{Si}}{C'_{ox} \left( v_{GS} - V_T - UTRA.v_{DS} \right)} \right]^{UEXP} \tag{7.80}$$

in which UCRIT and UTRA are fit parameters. The parameter UCRIT is denoted as the critical field for mobility reduction. A problem with this model is that the factor between the square brackets can become larger than one at some bias conditions, even with realistic values for the parameters in this model. This is physically not correct: the effective mobility will not increase due to the vertical field. In order to solve this problem, the effective mobility is set equal to  $\mu_0$  if the factor between the square brackets becomes larger than one. Hence, the mobility variation as a function of  $v_{GS}$  has a discontinuity. Such discontinuity can give rise to phantom harmonics or intermodulation products in numerical circuit simulations! The level 2 model of  $\mu_{eff}$  is not used anymore in modern MOS models of deep sub-micron devices.

#### 7.6.4.2 The mobility model of Liang et al.

The empirical model of Liang et al. [Liang 86] for the mobility dependence on the average normal field has been developed at the University of Berkeley. It is used for example in [Toh 88] and it forms the basis of the mobility model of the most recent versions of the BSIM model [BSIM 95]. With this model,  $\mu_{eff}$  is given by

$$\mu_{eff} = \frac{\mu_0}{1 + (E_{eff}/E_0)^{\nu}} \tag{7.81}$$

The values of  $\mu_0$ ,  $E_0$  and  $\nu$  are different for holes and electrons. They are given in Table 7.8. In

	$\mu_0 \ (cm^2/(V.s))$	$E_0 \ (MV/cm)$	$\nu$
electrons	670	0.67	1.6
holes	290	0.35	1.0

Table 7.8: Coefficients  $\mu_0$ ,  $E_0$  and  $\nu$  to be used in equation (7.81).

order to make this expression of practical use, the average vertical field  $E_{\it eff}$  is approximated by an expression in terms of the terminal voltages of the transistor. One finds in a semi-empirical way [Toh 88]

$$E_{eff} \approx \frac{v_{GS} + V_T}{6t_{or}} + \frac{V_T + V_a}{3t_{or}} \tag{7.82}$$

in which,  $V_a \approx 0.5V$  for typical  $n^+$  polysilicon gate devices [Toh 88].

The evaluation of the mobility model of equation (7.81) is time consuming since it contains a power function. In order to avoid this power function, a Taylor series expansion of equation (7.81) is used in the last BSIM versions, and the coefficients of this Taylor series are left to be determined using experimental data. The final expression for the effective mobility has been extended with a dependence on  $v_{SB}$ , and is given by [BSIM 95]

$$\mu_{eff} = \frac{\mu_0}{1 + \left[ UA \frac{v_{GS} + 2V_T}{t_{ox}} + UB \left( \frac{v_{GS} + 2V_T}{t_{ox}} \right)^2 \right] (1 - UC \cdot v_{SB})}$$
(7.83)

in which UA, UB and UC are to be determined experimentally. This equation shows again that mobility decreases when  $t_{ox}$  decreases. The equation shows some similarity with equation (7.78); but it has a higher-order term in  $v_{GS}$ .

# 7.6.5 Evaluation of nonlinearity coefficients

Having modeled the mobility reduction due to the normal field, it is interesting to investigate the influence of mobility reduction on the nonlinearity coefficients. The drain current equation from which the nonlinearity coefficients are derived is equation (7.70): this the most accurate equation we considered thus far. We will also discuss the impact on the nonlinearity coefficients of the different simplifications to the mobility model. It also has to be stressed that the nonlinearity coefficients computed in this section do not include yet other effects such as velocity saturation or the influence of a nonuniform doping. The influence of these effects is discussed in later sections.

We will only discuss the coefficients  $g_{mg}$ ,  $K_{2g_{mg}}$  and  $K_{3g_{mg}}$  that are proportional to derivatives of the drain current with respect to  $v_{GS}$  only, and the coefficients  $g_{ms}$ ,  $K_{2g_{ms}}$  and  $K_{3g_{ms}}$  that are proportional to derivatives with respect to  $v_{SB}$ . In the triode region, at low values of  $v_{DS}$  the nonlinearity coefficients that are proportional to derivatives with respect to  $v_{DB}$  are similar to  $g_{ms}$ ,  $K_{2g_{ms}}$  and  $K_{3g_{ms}}$ .

Next, we will compare the nonlinearity coefficients obtained with different mobility models. In this comparison we will consider nonlinearity coefficients with the source as a reference.

## 7.6.5.1 Triode region

First, the nonlinearity coefficients for the triode region are considered. The voltages are referred to the bulk, such that we will consider the nonlinearity coefficients that are defined in equation (7.4). The n-MOS transistor that we will consider in this section has the model parameters of the  $0.7\mu m$  process listed in Table 7.1. The value of  $\theta$  that is used in conjunction with equation (7.70) is different from the value given in Table 7.1. This table mentions mode parameters for the SPICE level 3 model, which models the mobility reduction more simply with equation (7.76). The new value of  $\theta$  has been obtained by performing a least square fit with the more accurate value of  $f_{\mu}$  of equation (7.69) onto the value of  $f_{\mu}$  from equation (7.76). This fit procedure is best performed in the triode region at a very low value of  $v_{DS}$  (typically 50mV) since this is the region of operation in which the value of  $\theta$  for the level 3 model has been extracted. The new value of  $\theta$  obtained in this way is  $0.05V^{-1}$ , which is lower than the value of to be used with the level 3 model.

Figure 7.7 depicts  $g_{mg} = \partial i_D/\partial v_{GB}$  as a function of  $V_{GB}$ . The other terminal voltage are  $V_{DB} = 1.45V$  and  $V_{SB} = 1V$ . If mobility reduction is neglected, which corresponds to  $\theta = 0V^{-1}$ , then the current is linearly dependent on  $v_{GB}$ , such that  $g_{mg}$  is constant. This has also been noticed in Section 7.4.2. However,  $g_{mg}$  decreases when  $\theta > 0$ , and it is no longer independent from  $v_{GB}$ . In other words, the mobility reduction causes a degradation of the (gates) transconductance, and it causes a nonlinear dependence of the current on  $v_{GB}$ .

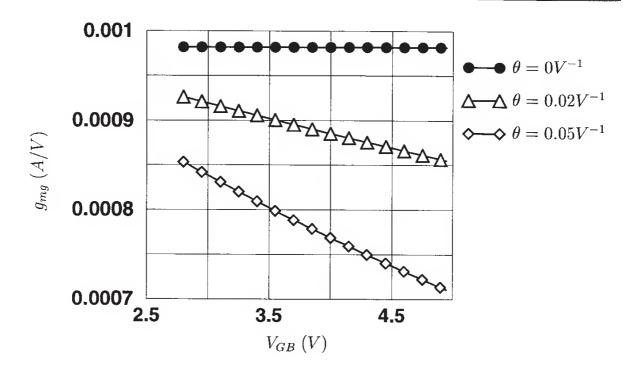


Figure 7.7:  $g_{mg}$  as a function of  $V_{GB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=1.45V$  and  $V_{SB}=1V$ .

Consider now the second-order nonlinearity coefficient  $K_{2g_{mg}}$ , which is the second-order derivative (multiplied with 1/2) of the drain current with respect to  $v_{GB}$ . Figure 7.8 depicts  $K_{2g_{mg}}$  as a function of  $V_{GB}$  in the triode region. Since  $g_{mg}$  is a decreasing function of  $V_{GB}$ ,  $K_{2g_{mg}}$  is negative for  $\theta > 0$ . The nonlinear dependence of the drain current on  $v_{GB}$  is very weak. Around  $V_{GB} = 3V$  the second-order normalized nonlinearity coefficient  $K'_{2g_{mg}}$  is around  $0.05V^{-1}$ .

Finally, the third-order nonlinearity coefficient  $K_{3g_{mg}}$  is shown as a function of  $V_{GB}$  in Figure 7.9. As for the second-order, we can conclude that the third-order nonlinearity coefficient is very small: the third-order normalized nonlinearity coefficient for  $\theta = 0.05V^{-1}$  is about  $5\times 10^{-3}V^{-2}$  for the  $V_{GB}$  range of Figure 7.9.

Next, the derivatives with respect to  $v_{SB}$  are considered. Figure 7.10 shows the small-signal parameter  $g_{ms} = -\partial i_D/\partial v_{SB}$  as a function of  $V_{SB}$ . The voltages  $V_{GB}$  and  $V_{DB}$  have been kept fixed to 4V and 2V, respectively. It is seen that  $g_{ms}$  does not change drastically as  $\theta$  becomes larger than zero.

Figures 7.11 and 7.12 depict the second-order nonlinearity coefficient  $K_{2g_{ms}}$  and  $K_{2g_{ms}}$ , respectively, as a function of  $V_{SB}$ .

Figure 7.12 shows that for  $\theta = 0.05V^{-1}$   $K_{3g_{ms}}$  goes through zero at  $V_{SB} = 1.4V$ . The fact that  $K_{3g_{ms}}$  can be made zero by applying an appropriate bulk bias, has been noticed for example in [Pat 90, Groen 94]. This is only possible due to the effect of mobility reduction. Indeed, with  $\theta = 0$   $K_{3g_{ms}}$  remains positive for all values of  $V_{SB}$ .

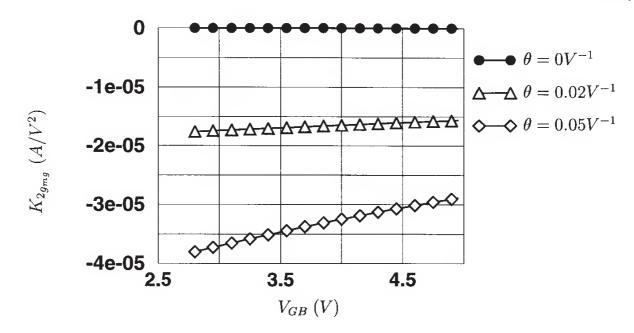


Figure 7.8:  $K_{2g_{mg}}$  as a function of  $V_{GB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=1.45V$  and  $V_{SB}=1V$ .

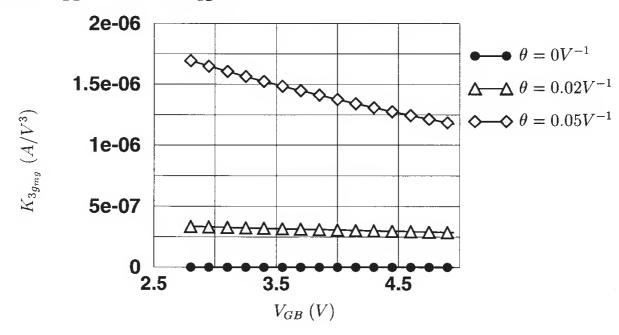


Figure 7.9:  $K_{3g_{mg}}$  as a function of  $V_{GB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=1.45V$  and  $V_{SB}=1V$ .

The zero-crossing of  $K_{3g_{ms}}$  at some value of  $V_{SB}$  forms the basis of the harmonic suppression technique described in [Pat 90]: when an AC voltage  $v_{ds}$  is applied over a MOS transistor with

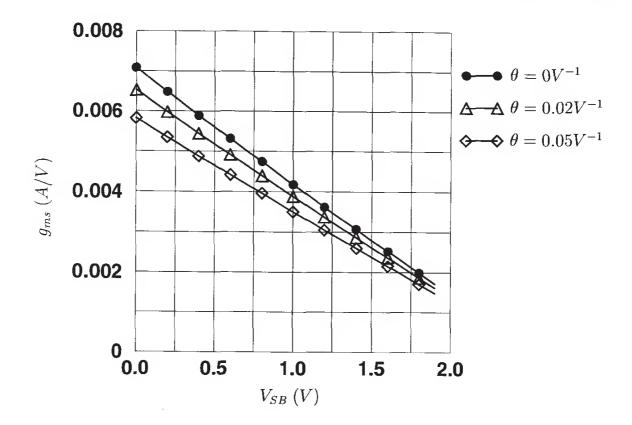


Figure 7.10:  $g_{ms}$  as a function of  $V_{SB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=2V$  and  $V_{GB}=4V$ .

 $V_{DS}=0V$ , then the second and third harmonic can be suppressed by applying a fraction of the AC voltage at the bulk and another fraction at the gate, as shown in Figure 7.13.

Comparison of nonlinearity coefficients obtained with different mobility models The function mobred defined in equation (7.73) is always of the form  $1/(1+f_{\mu})$ . In many models  $f_{\mu}$  is simplified to  $(v_{GS}-V_T)$  (see equation (7.76)). Compared to the more exact expression of  $f_{\mu}$  given in equation (7.69), this simplified expression does not contain any dependence on  $v_{DS}$ . Also, the dependence of mobility reduction on  $v_{SB}$  is wrong, as mentioned in Section 7.6.2. Nevertheless, the simple expression of  $f_{\mu}$  is widely used due to its simplicity. In addition it leads to a closed-form expression for  $v_{DSAT}$ . It is now investigated how much the nonlinearity coefficients obtained with the simple expression for  $f_{\mu}$  deviate from the nonlinearity coefficients obtained with the more exact expression for  $f_{\mu}$ . To this purpose, the nonlinearity coefficients have been computed for an n-MOS transistor of the  $0.7\mu m$  process with the two expressions for  $f_{\mu}$ . The value of  $\theta$  that needs to be used in conjunction with the simple expression for  $f_{\mu}$  is obtained from the set of SPICE level 3 model parameters. For the value of  $\theta$  that is to be used with the more exact expression of  $f_{\mu}$ , no measurement data are available. A realistic value of  $\theta$  can be obtained by fitting, as mentioned above. For a more complete comparison, however, it is interesting to

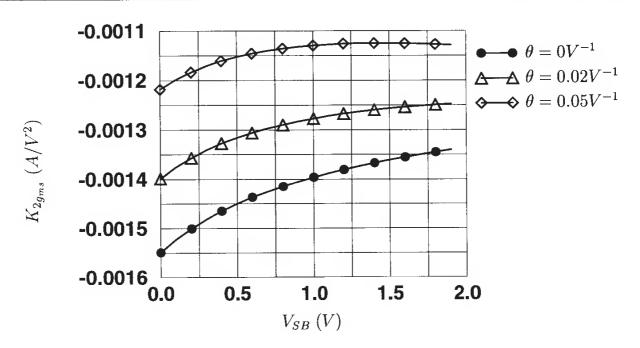


Figure 7.11:  $K_{2g_{ms}}$  as a function of  $V_{SB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=2V$  and  $V_{GB}=4V$ .

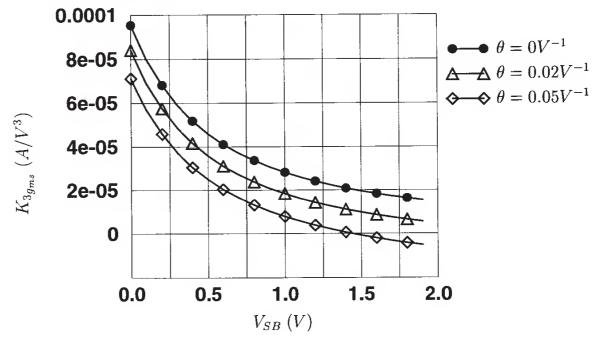


Figure 7.12:  $K_{3g_{ms}}$  as a function of  $V_{SB}$ , computed from the drain current model of equation (7.70) and with different values of  $\theta$  for an n-MOS transistor in the triode region with  $W=320\mu m$  and  $L=14\mu m$ . The model parameters of the  $0.7\mu m$  process are used except for  $\theta$ . Further  $V_{DB}=2V$  and  $V_{GB}=4V$ .

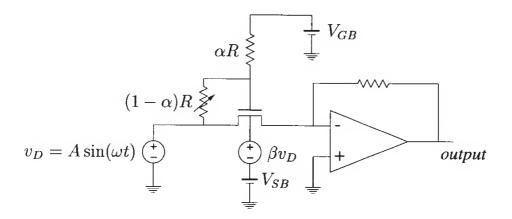


Figure 7.13: Measurement setup for the technique described in [Shou 93] to suppress the harmonics of the drain current of a transistor in the triode region. The parameters  $\alpha$  and  $\beta$  are numbers between 0 and 1.

compare the nonlinearity coefficients for several values of  $\theta$ . The comparison made in this section is limited to the nonlinearity coefficients that correspond to derivatives with respect to one single voltage. Nonlinearity coefficients that correspond to cross-derivatives are not considered here.

Figure 7.14 shows the error between the nonlinearity coefficients that have been computed with the simple mobility reduction model of equation (7.76) and the coefficients computed with the more exact model of equation (7.69). This simple model corresponds to the mobility reduction model of SPICE level 3, and hence the value of  $\theta$  that is to be used with this model can be taken from the list of SPICE level 3 parameters (Table 7.1). With the more exact model, three values of  $\theta$  have been considered: first the value that has been obtained by fitting as explained above. This value is  $0.05V^{-1}$ . Next, the same value as for the simple model  $(0.079V^{-1})$  is used, and finally, a value in between the two previous values is taken, namely  $0.065V^{-1}$ . The other model parameters have been taken from Table 7.1. For each coefficient the error has been referred to the value obtained with the simple model, and it is expressed in %. The nonlinearity coefficients are computed at one operating point in the triode region, namely for  $v_{DS} = 0.45V$ ,  $v_{GS} = 1.9V$  and  $v_{SB} = 1.3V$ . The transistor dimensions are  $W = 320\mu m$  and  $L = 14\mu m$ .

It is seen that for the value of the first-order derivatives, which are nothing but the small-signal parameters  $g_m$ ,  $g_o$  and  $g_{mb}$ , the deviation between the simple mobility reduction model and the more complicated one is smaller than 20%, and the error is minimal for  $\theta=0.05V^{-1}$ . One could conclude that for hand calculations on linearized circuits the use of the simplified mobility reduction model is justified. Although we mentioned in Section 7.6.2 that the dependence of the simplified mobility reduction model on  $v_{SB}$  is wrong, it is seen that the deviation of  $g_{mb}$  obtained with the more exact model and  $g_{mb}$  obtained with the simple mobility reduction model is small. This means that mobility reduction does not largely influence the value of  $g_{mb}$ .

Consider now the nonlinearity coefficients which are proportional to the higher-order derivatives. It is seen that the deviation between the simple mobility reduction model and the more exact one increases with the order of the derivative that is considered. Also, the value of  $\theta$  that

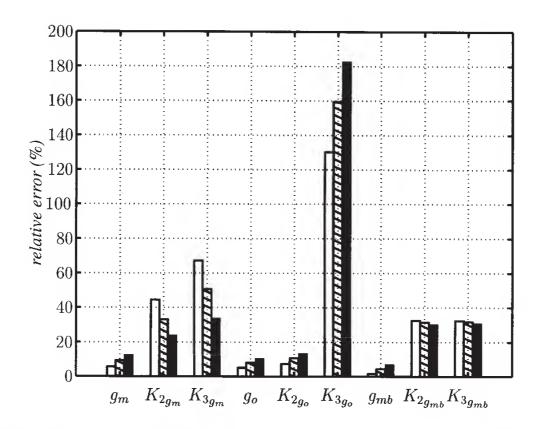


Figure 7.14: Relative error in % between the nonlinearity coefficients of a n-MOS transistor in the triode region ( $W = 320 \mu m$ ,  $L = 14 \mu m$ ,  $v_{DS} = 0.45 V$ ,  $v_{GS} = 1.9 V$ ,  $v_{SB} = 1.3 V$ ) computed with the simple mobility reduction model of equation (7.76) and the coefficients computed with the more exact model of equation (7.69). With the latter model three values of  $\theta$  have been considered:  $\theta = 0.05 V^{-1}$ , corresponding to the white bars,  $\theta = 0.065 V^{-1}$ , corresponding to the hatched bars, and  $\theta = 0.079 V^{-1}$ , corresponding to the black bars.

was found to give the smallest deviation for the current and its first-order derivatives, does not necessarily yield the smallest deviations for the higher-order derivatives.

It is seen that for the nonlinearity coefficients that are proportional to the second- and third-order derivatives with respect to  $v_{SB}$ , the error between the simple and the more complicated mobility reduction model remains smaller than 40%. In Section 7.6.2 it was mentioned that the dependence on  $v_{SB}$  of the mobility reduction is modeled wrong with the simple mobility reduction model of equation (7.77). The effect of this wrong modeling on the second- and third-order nonlinearity coefficients is limited to less than 40%. This can be explained by the fact that the influence of the variation of the depletion layer along the channel has a larger influence on the coefficients  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  than mobility reduction.

For the nonlinearity coefficients proportional to  $v_{GS}$  and  $v_{DS}$  the deviations can be much higher, as seen in Figure 7.14. For the nonlinearity coefficient  $K_{3g_o}$ , which is proportional to  $\partial^3 i_D/\partial v_{DS}^3$ , the error even exceeds 100%! Recall that with the simple drain current model of level 1, equation (7.25), where a uniform depletion layer is assumed and mobility reduction is neglected, the third-order derivative  $\partial^3 i_D/\partial v_{DS}^3$  is zero. When the variation of the depletion layer width along the channel is taken into account, then the drain current expression contains

 $\frac{3}{2}$  powers that contain  $v_{DS}$ , which give rise to a positive third-order derivative with respect to  $v_{DS}$ . With the introduction of the simple mobility reduction model this third-order derivative is just divided by a factor  $1 + \theta(v_{GS} - V_T)$  which is independent of  $v_{DS}$ . The deviation of more than 100% between  $K_{3g_o}$  obtained with the simple mobility reduction model and  $K_{3g_o}$  obtained with the more exact model, now shows that the dependence of mobility reduction on  $v_{DS}$  is important. The dependence on  $v_{DS}$  can be explained qualitatively as follows: mobility reduction is due to the vertical field, which is not purely determined by  $v_{GB}$ : it is clear that for  $v_{DS} > 0$  the vertical field at the drain end of the channel is different from the field at the source end. In other words, the vertical field, and hence the mobility reduction varies along the channel, which yields a dependence on  $v_{DS}$ .

#### 7.6.5.2 Saturation region

When the more exact mobility reduction model of equation (7.69) is used, then  $v_{DSAT}$  cannot be computed explicitly, as mentioned in Section 7.6.3. Nevertheless, explicit expressions can be obtained for the derivatives in terms of the terminal voltages and  $v_{DSAT}$ . In this section a comparison is made between the nonlinearity coefficients for a MOS transistor in saturation, computed with the more exact mobility reduction model and with the simple model of equation (7.76). Derivatives with respect to  $v_{DS}$  are not considered, since these are primarily determined by effects such as channel-length modulation, as will be discussed in Section 7.11.

Figure 7.15 shows a comparison between the simple and the more exact mobility reduction model for the coefficients  $g_m$ ,  $K_{2g_m}$ ,  $K_{3g_m}$ ,  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$ .

In the saturation region the dependence of the coefficients  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$  is not only through  $v_{GS}$  but also through  $v_{DSAT}$ : apart from channel-length modulation effects, which are not considered in this section, the expression of the drain current in saturation is the same as for the triode region but with  $v_{DS}$  replaced by  $v_{DSAT}$ , which in turn depends on  $v_{GS}$  and  $v_{SB}$ . The same holds for the dependence of the coefficients  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$ .

It is seen that just as in the triode region the deviation between the simple mobility reduction model and the more exact model increases with the order of the derivatives.

# 7.7 Velocity saturation

All models for the drain current derived so far are based upon the assumption that the drift velocity of carriers is linearly proportional to the longitudinal electric field. The constant of proportionality is the mobility, as stated in equation (7.15). This assumption is correct as long as the drift velocity is small compared to the thermal velocity of carriers  $v_{sat}$ , which is about  $10^7 cm/s$  for silicon at room temperature. As the drift velocity approaches the thermal velocity, its field dependence will begin to depart from the linear relationship: the velocity will not increase much anymore. This saturation effect is referred to as **velocity saturation** [Sze 85, Mull 86]. Since in that situation the carriers attain energies above the ambient thermal energy, they are often characterized as **hot carriers**. Also, the thermal velocity  $v_{sat}$  is often referred to as the **saturation velocity**. Different values have been reported for the value of  $v_{sat}$  [Coen 80, Coop 81, Sze 85, Gar 87a]

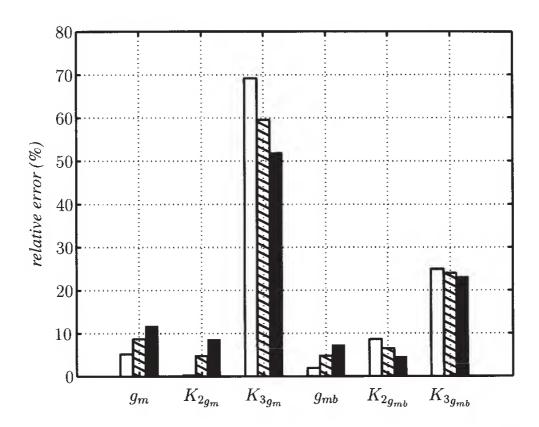


Figure 7.15: Relative error in % between the nonlinearity coefficients of a n-MOS transistor in the saturation region ( $W=320\mu m$ ,  $L=14\mu m$ ,  $v_{DS}=1.45V$ ,  $v_{GS}=1.9V$ ,  $v_{SB}=1.3V$ ) obtained with the simple mobility reduction model and the coefficients obtained with the more exact model. With the latter model three values of  $\theta$  have been considered:  $\theta=0.05V^{-1}$ , corresponding to the white bars,  $\theta=0.065V^{-1}$ , corresponding to the hatched bars, and  $\theta=0.079V^{-1}$ , corresponding to the black bars.

ranging from  $5 \times 10^4 m/s$  to  $1.2 \times 10^5 m/s$  for electrons and from  $4 \times 10^4 m/s$  to  $8 \times 10^4 m/s$  for holes. The dependences on gate-oxide thickness, doping concentration and other processing conditions have not been established [Gar 87a, Koh 89] although  $v_{sat}$  appears to be relatively independent of the gate field. Values for  $v_{sat}$  that are used in MOS models can vary over an even larger range, since these values are obtained by fitting. In this way,  $v_{sat}$  is rather a fitting parameter than a physical parameter.

Intuitively, we can say that the characteristics of short-channel devices will be more influenced by velocity saturation than long-channel devices. If we assume for simplicity that the lateral electric field is constant along the channel and given by  $v_{DS}/L$ , then it is clear that for the same  $v_{DS}$ , the electric field in a short-channel device will be higher than in a long-channel device. As a result, velocity saturation will occur at lower drain-source voltages for a short-channel device. The real situation is more complicated than the above reasoning, which only gives a qualitative idea.

Velocity saturation is a second mechanism that decreases the mobility of carriers. Whereas the previous section discussed the mobility reduction due to the presence of the normal electric field, this section deals with mobility reduction due to the lateral electric field. Many authors refer

to mobility reduction only to address the influence of the normal field, whereas the influence of the lateral field is just referred to as velocity saturation.

Before deriving a model for the drain current including the effect of velocity saturation, different models that relate the velocity to the electric field will be considered.

### 7.7.1 Velocity-field models

An empirical model for velocity saturation relates the drift velocity v of the carriers and the longitudinal electric field  $E_x$  as follows [Sodi 84, Sze 85, Mull 86, Matt 96]

$$v = \frac{\mu_{eff} E_x}{\left[1 + \left(\frac{E_x}{E_c}\right)^n\right]^{1/n}} \tag{7.84}$$

Here  $\mu_{eff}$  is the effective mobility, given by one of the models of Section 7.6,  $E_c$  is the *critical* electric field, given by

$$E_c = \frac{v_{sat}}{\mu_{eff}} = \frac{v_{sat}}{\mu_0} (1 + f_\mu) \tag{7.85}$$

and n is a fitting parameter between 1 and 2. According to several authors [Sodi 84, Sze 85], the best value for electrons is n = 2, whereas n = 1 gives the best results for holes.

In order to have a first quantitative idea about the effect of velocity saturation, we evaluate  $E_c$  for the  $0.7\mu m$  reference process of Table 7.1. Assume that the mobility is not reduced, such that  $\mu_{eff}=\mu_0=0.047m^2/(V.s)$ . The velocity  $v_{sat}$  is taken equal to  $10^5m/s$ . Then  $E_c$  is found to be 2.13MV/m. Assume that a transistor is biased in the triode region. The average lateral electrical field can be estimated as  $v_{DS}/L$ . If  $v_{DS}=0.7V$  and  $L=0.7\mu m$  then the lateral field equals 1MV/m and the ratio  $E/E_c$  equals 0.47.

Equation (7.84) can be used in equation (7.14) to derive an expression for the drain current. However, if  $n \neq 1$ , then this leads to difficulties to obtain an analytical expression for the drain current. This is shown in Appendix G. In this appendix several approximations are suggested to obtain a closed-form expression for the current. This issue will be discussed further in this section.

In order to obtain a simple transistor model various simplified expressions are used to model velocity saturation. In Figure 7.16 three models are shown for the relationship between the electron velocity and the electric field. A very accurate model for electrons, referred to as model 1 in Figure 7.16 is the model of equation (7.84) with n=2. A second velocity-field model, in Figure 7.16 referred to as model 2, is derived from equation (7.84) by setting the fitting parameter n equal to 1. Then the relationship between velocity and field reduces to

$$v = \frac{\mu_{eff} E_x}{1 + \frac{E_x}{E_c}} \tag{7.86}$$

With this model it is seen that the velocity is underestimated compared to model 1. In addition, it is seen that the velocity deviates earlier from the linear relationship than with model 1. This is

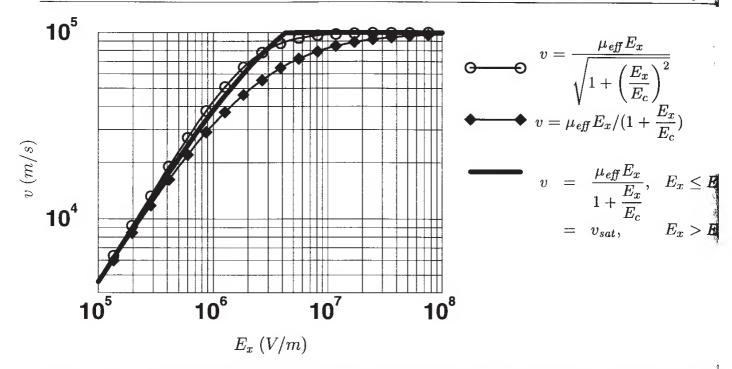


Figure 7.16: Comparison of three commonly used velocity-field models for velocity saturation [Sodi 84].

explained by the fact that in model 1 the term  $(E_x/E_c)^2$  is much smaller than the term  $(E_x/E_c)^2$  in model 2 as long as  $E_x \ll E_c$ . In fact, velocity saturation occurs much more abruptly with model 1 than with model 2.

A third model, referred to as model 3 in Figure 7.16, has been proposed in [Sodi 84]. It is a piecewise model, given by

$$v = \frac{\mu_{eff} E_x}{1 + \frac{E_x}{E_c}}, \quad E_x \le E_c$$

$$= v_{sat}, \quad E_x > E_c$$

$$(7.87)$$

where  $E_c$  is now given by

$$E_c = 2v_{sat}/\mu_{eff} \tag{7.88}$$

which is a factor two larger than the value of the critical field used in models 1 and 2. It is seen that with the piecewise approximation of model 3 the velocity of carriers is equal to  $v_{sat}$  at the critical field, whereas in the two other models the electric field  $E_x$  must be much larger than the critical field  $E_c$  for the velocity to approach  $v_{sat}$ . Compared to model 2, the piecewise model 3 has a steeper slope at high fields and it has a better correspondence with model 1. As a result model 3 is widely used in many modern transistor models [Toh 88, Moon 91, BSIM 95].

# 7.7.2 Drain current in the triode region

In Appendix G expressions for the drain current in the triode region are derived using the models 1, 2 and 3. These expressions are now briefly discussed.

#### 7.7.2.1 Drain current with the simple velocity-field models

For both model 2 and 3, the drain current is found to be (see Appendix G)

$$i_{D} = \frac{W}{L} \frac{\mu_{eff} C'_{ox}}{\left(1 + \frac{(v_{DB} - v_{SB})}{LE_{c}}\right)} \left\{ (v_{GB} - V_{FB} - \phi) (v_{DB} - v_{SB}) - \frac{1}{2} (v_{DB}^{2} - v_{SB}^{2}) - \frac{2}{3} \gamma \left[ (\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$

$$(7.89)$$

The value of the critical electric field is either given by  $v_{sat}/\mu_{eff}$  (model 2) or  $2v_{sat}/\mu_{eff}$  (model 3). Expression (7.89) is the extension of equations (7.37) and (7.70) that take into account a nonuniform depletion layer. The effective mobility  $\mu_{eff}$  is equal to  $\mu_0/(1+\theta f_\mu)$  and  $f_\mu$  is given in equation (7.68).

#### 7.7.2.2 The functions large, mobred and hot

It is seen that equation (7.89) is the product of three factors: the first factor is the drain current that has been derived in Section 7.4.2 without taking into account mobility reduction and velocity saturation. We denote this factor as the function *large*:

$$large (v_{GB}, v_{DB}, v_{SB}) = \frac{\mu_0 C'_{ox} W}{L} \left\{ (v_{GB} - V_{FB} - \phi) (v_{DB} - v_{SB}) - \frac{1}{2} (v_{DB}^2 - v_{SB}^2) - \frac{2}{3} \gamma \left[ (\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(7.90)

The second factor models the mobility reduction due to the normal field and is denoted as the function mobred. This function has been defined in equation (7.73). It is clear that the function mobred ( $v_{GB}$ ,  $v_{DB}$ ,  $v_{SB}$ ) is never larger than one.

The third factor models velocity saturation, and is written as  $hot(v_{GB}, v_{DB}, v_{SB})$ . Using the velocity-field models 2 and 3, the function hot is given by

$$hot(v_{GB}, v_{DB}, v_{SB}) = 1 / \left(1 + \frac{(v_{DB} - v_{SB})}{LE_c}\right)$$
 (7.91)

and it should never be larger than one. This function depends on  $v_{GB}$  through the critical electric field  $E_c$ : this is equal to  $v_{sat}/\mu_{eff}$  (velocity-field model 2) or  $2v_{sat}/\mu_{eff}$  (model 3), and  $\mu_{eff}$  depends on  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$  since it is equal to  $\mu_0 \cdot mobred(v_{GB}, v_{DB}, v_{SB})$ .

The resulting drain current can then be written as

$$i_D = mobred(v_{GB}, v_{DB}, v_{SB}) \cdot hot(v_{GB}, v_{DB}, v_{SB}) \cdot large(v_{GB}, v_{DB}, v_{SB})$$

$$(7.92)$$

Many designers prefer to reason with voltages referred to the source. Then the drain current can be rewritten as

$$i_D = mobred(v_{GS}, v_{DS}, v_{SB}) \cdot hot(v_{GS}, v_{DS}, v_{SB}) \cdot large(v_{GS}, v_{DS}, v_{SB})$$

$$(7.93)$$

with the functions mobred, hot and large given by

$$large (v_{GS}, v_{DS}, v_{SB}) = \frac{\mu_0 C'_{ox} W}{L} \left\{ (v_{GS} - V_{FB} - \phi) v_{DS} - \frac{1}{2} v_{DS}^2 - \frac{2}{3} \gamma \left[ (\phi + v_{SB} + v_{DS})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(7.94)

$$mobred(v_{GS}, v_{DS}, v_{SB}) = 1/(1 + \theta f_{\mu}(v_{GS}, v_{DS}, v_{SB}))$$
 (7.95)

$$hot(v_{GS}, v_{DS}, v_{SB}) = 1 / \left(1 + \frac{v_{DS}}{LE_c}\right)$$
 (7.96)

In previous sections, we saw other models for the functions *large* and *mobred*. As a result, many possibilities can be combined in order to obtain a drain current model. As an example we consider the functions *large*, *mobred* and *hot* in a simplified representation of the BSIM3 model. The function *large* in this model is given by

$$large\left(v_{GS}, v_{DS}, v_{SB}\right) = \frac{W}{L} \mu_0 C'_{ox} \left[ \left(v_{GS} - V_T\right) v_{DS} - \frac{1}{2} a v_{DS}^2 \right]$$
(7.97)

in which parameter a depends on the bulk bias (see Section 7.4.3). The function mobred for the BSIM model is given by (see Section 7.6.4.2)

$$mobred (v_{GS}, v_{DS}, v_{SB}) = \frac{1}{1 + \left[ UA \frac{v_{GS} + 2V_T}{t_{ox}} + UB \left( \frac{v_{GS} + 2V_T}{t_{ox}} \right)^2 \right] (1 - UC \cdot v_{SB})}$$
(7.98)

Velocity saturation in the BSIM model is modeled with velocity-field model 3, such that the function hot is given by equation (7.96) with  $E_c = 2v_{sat}/\mu_{eff}$ .

## 7.7.2.3 Merging of the functions mobred and hot

It is seen that both the functions mobred and hot are of the form 1/(1+x), in which x depends on model parameters and on the terminal voltages. In several MOS models [Merck 72, White 80, Gar 87a] it is assumed that the effect of mobility reduction due to the normal field and velocity saturation are so small that they can be merged into one factor. Indeed, if the product of mobred and hot is represented as

$$mobred \cdot hot = \frac{1}{1+x} \cdot \frac{1}{1+y} \tag{7.99}$$

and if both x and y are much smaller than one, then

$$\frac{1}{1+x} \cdot \frac{1}{1+y} = \frac{1}{1+x+y+xy} \approx \frac{1}{1+x+y} \tag{7.100}$$

This approach is often used in hand calculations with very simple models for the functions mobred and hot. When the effective mobility is written as  $\mu_{eff} = \mu_0/(1 + \theta(v_{GS} - V_T))$  (see equation (7.77)) and velocity saturation is modeled as in equation (7.96) then, using the approximation made in equation (7.100), one obtains for the drain current in the triode region

$$i_D = \frac{large}{1 + \theta(v_{GS} - V_T) + \frac{v_{DS}}{LE_c}}$$

$$(7.101)$$

Using the function large of equation (7.97) one obtains

$$i_D = \frac{W}{L} \mu_0 C'_{ox} \frac{(v_{GS} - V_T) v_{DS} - \frac{1}{2} a v_{DS}^2}{1 + \theta(v_{GS} - V_T) + \frac{v_{DS}}{LE_c}}$$
(7.102)

Let us consider the error on the drain current that is made by the approximation of equation (7.100). With the model parameters of the  $0.7\mu m$  process we have  $V_T=0.75V$ ,  $\mu_0=0.047m^2/(V.s)$ ,  $\theta=0.079V^{-1}$  and  $v_{sat}=1.94\times10^5m/s$ . If  $v_{GS}-V_T=0.2V$  and  $v_{DS}=0.1V$ , then the factor  $\theta(v_{GS}-V_T)$  equals 0.0158, which is considerably smaller than 1. The critical electric field, given by  $v_{sat}/\mu_{eff}$  equals 4.19MV/m. Hence, for a transistor with a channel length of  $0.7\mu m$  the term  $v_{DS}/(LE_c)$  equals 0.034, which is much smaller than one as well. In this situation, the error made by merging the two factors into one term is 0.05%. When  $v_{GS}-V_T$  equals 1V and  $v_{DS}$  is 0.5V, then the factor  $\theta(v_{GS}-V_T)$  equals 0.079 and  $v_{DS}/(LE_c)$  equals 0.17. The error made by merging the two factors is 1%. The error on the nonlinearity coefficients can be larger. This is further discussed in Section 7.7.4.

An interesting observation is that for this technology velocity saturation has a larger influence on the drain current than mobility reduction due to the normal field. Indeed,  $\theta(v_{GS}-V_T)$  is larger than  $v_{DS}/(LE_c)$  for the two considered bias points. This is usually the case for submicron technologies.

### 7.7.2.4 Drain current with the more accurate velocity-field model

In Section 7.7.1 it has been pointed out that the velocity-field model 1 is more accurate for electrons than the models 2 and 3. The disadvantage of using the more accurate velocity-field model is that it is impossible to obtain a closed-form expression for the drain current since the integral given in equation (7.19) cannot be computed analytically. In Appendix G an approximate expression is derived for the drain current. This is given by

$$i_D = mobred(v_{GB}, v_{DB}, v_{SB}) \cdot hot(v_{GB}, v_{DB}, v_{SB}) \cdot large(v_{GB}, v_{DB}, v_{SB})$$
(7.103)

with the function hot now given by

$$hot (v_{GB}, v_{DB}, v_{SB}) = \frac{1}{\sqrt{1 + \left(\frac{v_{DB} - v_{SB}}{LE_c}\right)^2}}$$
(7.104)

Accurate expressions for the functions *large* and *mobred* are given in equations (7.90) and (7.95), respectively. This yields the final expression of the drain current

$$i_{D} = \frac{\mu_{0}C'_{ox}W}{L} \left\{ (v_{GB} - V_{FB} - \phi) (v_{DB} - v_{SB}) - \frac{1}{2} (v_{DB}^{2} - v_{SB}^{2}) - \frac{2}{3}\gamma \left[ (\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\} \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \frac{1}{\sqrt{1 + \left(\frac{v_{DB} - v_{SB}}{LE_{c}}\right)^{2}}}$$
(7.105)

This accurate expression will form the basis for the evaluation of nonlinearity coefficients in Section 7.7.4.

Equation (7.103) is also used in [Grot 84]. In this reference, this equation has also been compared to the more exact value of the drain current obtained with numerical integration. The error made with equation (7.103) is very small.

The function hot as given in equation (7.104) does not change when  $v_{SB}$  and  $v_{DB}$  are interchanged. From Section 7.6.2 we recall that this is also true for the function mobred of equation (7.73). On the other hand, when the role of  $v_{DB}$  and  $v_{SB}$  is interchanged in the function large, then this function changes sign. When the three functions are combined to form the drain current, then it is seen that the latter changes sign as  $v_{DB}$  and  $v_{SB}$  are swapped. This is important to model the drain current and its derivatives around  $v_{DS} = 0V$ , as will be explained in the next section.

### **7.7.2.5** Modelling around $v_{DS} = 0V$

Consider a MOS transistor that is biased in strong inversion with  $v_{DS} = 0V$ . Assume now that a sinusoidal voltage is applied over the transistor:

$$v_{DS} = A\sin\left(\omega t\right) \tag{7.106}$$

This situation is depicted in Figure 7.17. In this way,  $v_{DS}$  is negative during half of the period. If  $v_{DS}$  is negative and velocity saturation is modeled with the function hot given by  $1/(1+v_{DS}/(LE_c))$ , then hot becomes larger than one for negative  $v_{DS}$  values. This is physically impossible, since it would imply that the drain current increases due to velocity saturation. Clearly, the function hot which was seen in the previous section to be symmetric with respect to source and drain when the accurate velocity-field model is used, is no longer symmetric when the simple velocity-field models 2 or 3 are used. This problem could be solved by taking the absolute value of  $v_{DS}$ . Then the function hot becomes

$$hot\left(v_{GS}, v_{DS}, v_{SB}\right) = 1 \left/ \left(1 + \left| \frac{v_{DS}}{LE_c} \right| \right) \right. \tag{7.107}$$

However, this function will give problems when derivatives or nonlinearity coefficients need to be evaluated. The derivative of the drain current with respect to  $v_{DS}$  contains the derivative of

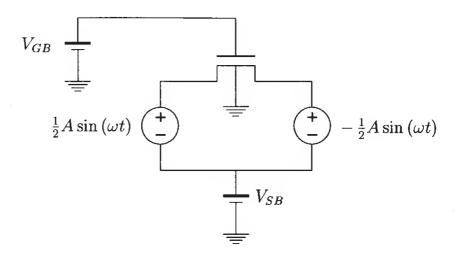


Figure 7.17: A MOS transistor in the triode region with  $V_{DS}=0V$  and with a symmetrical sinusoidal input excitation.

the function hot with respect to  $v_{DS}$ . From equation (7.107) we find that, if  $E_c$  is assumed for simplicity to be independent of  $v_{DS}$ , this derivative is given by

$$\frac{dhot}{dv_{DS}} = \frac{1}{\left(1 + \frac{|v_{DS}|}{LE_c}\right)^2} \cdot \frac{1}{LE_c} \cdot \frac{|v_{DS}|}{v_{DS}}$$
(7.108)

This derivative does not exist for  $v_{DS} = 0V$  since the last factor is undefined for zero  $v_{DS}$ . Hence, the velocity-field models 2 and 3 are not only less accurate than model 1 to model velocity saturation, but they also give problems to compute derivatives at  $v_{DS} = 0V$ .

# 7.7.3 Drain current in the saturation region

The drain current in the saturation regime can be determined in the same way as we did before in Sections 7.5 and 7.6.3: first,  $v_{DSAT}$  is determined from solving  $\partial i_D/\partial v_{DS}=0$  for  $v_{DS}$ , after which  $v_{DSAT}$  is substituted into the drain current expression for the triode region. When velocity saturation is neglected, as we did in previous sections, then the condition  $\partial i_D/\partial v_{DS}=0$  corresponds to the physically impossible assumption that the inversion layer charge is zero, such that the drift velocity should be infinite in order to allow for a finite, nonzero drain current. This assumption can now be corrected with the inclusion of velocity saturation into the transistor model.

# 7.7.3.1 Drain current in saturation with the simple velocity-field models

As we saw in previous sections, the expression for the drain current in saturation starts from the expression of the current in the triode region. Let us describe the velocity-field relationship with

model 2 (equation (7.86)). If the depletion layer is assumed to vary linearly along the channel, then the drain current in the triode region is given by (see Appendix G)

$$i_D = \frac{W}{L} \frac{\mu_{eff} C'_{ox} \left[ (v_{GS} - V_T) v_{DS} - \frac{1}{2} a v_{DS}^2 \right]}{1 + v_{DS} / (LE_c)}$$
(7.109)

with  $E_c$  given by  $v_{sat}/\mu_{eff}$ . Assume now for simplicity that  $\mu_{eff}$  is independent of  $v_{DS}$ . When  $v_{DSAT}$  is determined from equation (7.109) by solving  $\partial i_D/\partial v_{DS}=0$  for  $v_{DS}$ , then one easily finds [Tsiv 88]

$$v_{DSAT} = LE_c \left( \sqrt{1 + \frac{2(v_{GS} - V_T)}{aLE_c}} - 1 \right)$$
 (7.110)

This value is smaller than the value of  $(v_{GS} - V_T)/a$  that is obtained in the absence of velocity saturation effects (see equation (7.61)). If  $LE_c$  goes to infinity, which corresponds to the absence of velocity saturation, then  $v_{DSAT}$  again approaches  $(v_{GS} - V_T)/a$ .

Let us now consider the value of the inversion layer charge at the drain end when  $v_{DS}$  equals the value of  $v_{DSAT}$  given in equation (7.110). Using equation (7.46), this charge equals

$$Q'_{I_{drain}} = -C'_{ox} \left( v_{GS} - V_T - a \, v_{DSAT} \right) \tag{7.111}$$

This charge is larger than zero since  $v_{DSAT} < (v_{GS} - V_T)/a$ . This is a more realistic situation, compared to the situation where velocity saturation is neglected. Indeed, without velocity saturation the onset of saturation is reached when  $Q'_{I_{drain}} = 0$  which means that the electrons have to drift with an infinite velocity near the drain end of the channel.

Instead of deriving  $v_{DSAT}$  in the classical way, an alternative approach can be followed, which leads to exactly the same value of  $v_{DSAT}$ . Recall that the current can be written as (see equation (7.14))

$$i_D = W. (-Q_I'(x)) . v(x)$$
 (7.112)

At the onset of saturation it is assumed that the velocity of carriers at the drain end is equal to  $v_{sat}$ . With this assumption, the drain current in the saturation regime (neglecting the dependence on  $v_{DS}$  for example by channel-length modulation) is also given by

$$i_{DSAT} = W. \left( -Q'_{I_{drain}} \right) . v_{sat} \tag{7.113}$$

Using equations (7.111) and (7.110) one can easily compute that this expression is identical to equation (7.109) evaluated at  $v_{DSAT}$ .

As  $v_{DS}$  is increased above  $v_{DSAT}$ , the channel consists of two parts: one part, which starts from the source in which the velocity is field-dependent, and one part, close to the drain, where velocity saturates. In the latter part, the electric field is very high, but not infinite, as was assumed implicitly in the derivations of Section 7.5. In fact, the pinch-off condition that the mobile charge becomes zero at the drain end, never occurs, even for a long-channel transistor.

The assumption that the velocity of carriers approaches  $v_{sat}$  at the drain end when saturation starts, is also used to determine  $v_{DSAT}$  when velocity saturation is modeled with the piecewise velocity-field model. With this model, the expression of  $v_{DSAT}$  is somewhat simpler than with model 2:

$$v_{DSAT} = \frac{E_c L (v_{GS} - V_T)}{a E_c L + v_{GS} - V_T}$$
(7.114)

This expression has again been obtained by assuming that  $\mu_{eff}$  is independent of  $v_{DS}$ . The value of the saturation voltage given in equation (7.114) is somewhat smaller than with model 2. Also, it does not correspond anymore to  $\partial i_D/\partial v_{DS}=0$ . This is due to the piecewise approximation in the velocity-field model. Again, this expression reduces to  $(v_{GS}-V_T)/a$  for transistors with a long channel. On the other hand, if  $(v_{GS}-V_T)/a$  is kept constant and L is scaled down, then  $t_{ox}$  will be smaller as well, according to the scaling laws. As a result,  $\theta$  increases. Consequently,  $\mu_{eff}$  decreases, which causes a slight increase of  $E_c$ . However, the net result is that the product  $LE_c$  will decrease. This means that  $v_{DSAT}$  will decrease. In other words, the  $v_{DS}$  range over which the transistor operates in the triode region becomes smaller when the channel length decreases.

When the value of  $v_{DSAT}$  given in equation (7.114) is substituted into equation (7.109), then the current in the saturation region becomes

$$i_{DSAT} = W v_{sat} C'_{ox} (v_{GS} - V_T - a v_{DSAT})$$
 (7.115)

This model, of course, does not take into account yet effects such as channel-length modulation that determine the output conductance. However, the model of the output conductance that will be given in Section 7.11 is based upon equation (7.115).

### 7.7.3.2 Merging of the functions mobred and hot

In Section 7.7.2.3 a drain current model has been presented in which the product mobred.hot was merged into one single function. In [Gar 87a] this model is used in conjunction with the piecewise velocity-field model of equation (7.87). In this reference, the saturation voltage  $v_{DSAT}$  is computed with the assumption that at the onset of saturation, the velocity of carriers is equal to  $v_{sat}$ . The value of  $v_{DSAT}$  found in this way is given by

$$v_{DSAT} = \frac{\left[1 + \theta \left(v_{GS} - V_T\right)\right] \left(v_{GS} - V_T\right)}{a \left[1 + \left(\theta + \frac{\mu_0}{2av_{sat}L}\right) \left(v_{GS} - V_T\right)\right]}$$
(7.116)

in which a is defined in Section 7.4.3. The corresponding drain current is given by

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{a L} \frac{\left(v_{GS} - V_T\right)^2}{1 + \left(\theta + \frac{\mu_0}{2av_{sat}L}\right) \left(v_{GS} - V_T\right)}$$
(7.117)

In conjunction with velocity-field model 2, equation (7.117) is often simplified to [Hspi 96]

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{a L} \frac{\left(v_{GS} - V_T\right)^2}{1 + \left(\theta + \frac{\mu_0}{v_{sat} L}\right) \left(v_{GS} - V_T\right)}$$
(7.118)

Introducing the function mobhot

$$mobhot(v_{GS}, v_{SB}) = \frac{1}{1 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(v_{GS} - V_T)}$$
(7.119)

the drain current given in equation (7.118) becomes

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{a L} (v_{GS} - V_T)^2 \cdot mobhot(v_{GS}, v_{SB})$$
 (7.120)

The function *mobhot* indicates how much the drain current is reduced by the combination of mobility reduction due to a vertical field and velocity saturation. This function is always smaller than 1.

Equation (7.118) is widely used for hand calculations with submicron devices. As an example, consider a  $0.7\mu m$  transistor fabricated with the  $0.7\mu m$  technology the parameters of which are listed in Table 7.1 and with  $v_{GS}=1.9V$  and  $v_{SB}=1.3V$ . Then the value of the function mobhot is 0.763.

#### 7.7.3.3 Drain current in saturation with the more accurate velocity-field model

When the more accurate velocity-field model 1 (equation (7.84)) is used, then  $v_{DSAT}$  is determined starting from the drain current expression of equation (7.103). If the effective mobility is assumed to be independent of  $v_{DS}$  and the charge of the depletion layer is assumed to vary linearly along the channel, then setting  $\partial i_D/\partial v_{DS}$  equal to zero and solving for  $v_{DS}$  requires the determination of a third-order polynomial in  $v_{DS}$ . This leads to a very complicated expression.

If more accurate expressions for the effective mobility are used, such that it has a dependence on  $v_{DS}$  and if the variation of the depletion layer along the channel is modeled more accurately, then setting  $\partial i_D/\partial v_{DS}$  to zero leads to a complicated expression from which  $v_{DS}$  cannot be solved explicitly anymore. Instead,  $v_{DSAT}$  must be determined by iteration. This can be performed efficiently: using for example expression (7.114) as a starting point, a good initial approximation is already obtained, and very few extra iterations are required. Once  $v_{DSAT}$  is known, the drain current in saturation can be obtained by substituting the value of  $v_{DSAT}$  into the expression for the drain current in the triode region (see equation (7.103)).

Since  $v_{DSAT}$  cannot be computed explicitly, it is impossible to express the drain current in terms of the terminal voltages of the transistor:  $v_{DSAT}$  cannot be eliminated from the expression of the current. Consequently, the derivatives of the current with respect to a terminal voltage will also contain  $v_{DSAT}$ . When computing the derivatives of the current with respect to  $v_{GS}$  or  $v_{SB}$  one must take into account that  $v_{DSAT}$  is a function of these two voltages. In Appendix F

it is explained how these derivatives can be computed. These derivatives will be used when the nonlinearity coefficients are obtained.

The drain current in terms of  $v_{GS}$ ,  $v_{SB}$  and  $v_{DSAT}$  is found from equations (7.94), (7.95) and (7.104) to be

$$i_{DSAT} = \frac{\mu_0 C'_{ox} W}{L} \left\{ (v_{GS} - V_{FB} - \phi) v_{DSAT} - \frac{1}{2} v_{DSAT}^2 - \frac{1}{2} v_{DSAT}^2 - \frac{1}{2} \gamma \left[ (\phi + v_{SB} + v_{DSAT})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\} \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \frac{1}{\sqrt{1 + \left(\frac{v_{DSAT}}{LE_c}\right)^2}}$$
(7.121)

It should be noted that the dependence on  $v_{DS}$  through effects such as channel-length modulation is not included here.

### 7.7.4 Evaluation of nonlinearity coefficients

With the effects that have been studied thus far in this chapter, we have obtained the accurate drain current expressions of equation (7.105) and (7.121). These equations model different effects that influence the nonlinearity coefficients:

- the nonlinear relationship between the drain current and the terminal voltages. This relation is also nonlinear in the simple SPICE level 1 model: for example, the drain current in the triode region depends on  $v_{DS}^2$ , which gives rise to a nonzero second-order nonlinearity coefficient  $K_{2g_0}$ .
- the variation of the depletion layer width along the channel. This gives rise to terms with  $\frac{3}{2}$  powers in the drain current expression. These terms give rise to other values for the nonlinearity coefficients than obtained with the level 1 model.
- mobility reduction.
- velocity saturation.

In Section 7.8 and in subsequent sections we will discuss some additional effects that influence the drain current and the nonlinearity coefficients.

Some nonlinearity coefficients are primarily determined by only one of the above effects, whereas other ones can be largely influenced by more than one of these effects. It is instructive to know which effect primarily determines a nonlinearity coefficient. This can be accomplished with the approach that has been discussed in Section 3.5 to determine the dominant terms of the expression of a nonlinearity coefficient.

When we derive the nonlinearity coefficients from equation (7.105) and (7.121) then we will assume that the critical field  $E_c$  is independent of the terminal voltages. This is not exact, since  $E_c = v_{sat}/\mu_{eff}$  and  $\mu_{eff}$  depends on the terminal voltages. Nevertheless, the error on the nonlinearity coefficients due to this simplification remains smaller than 1%.

The model parameters that will be used are the ones of the  $0.7\mu m$  process, listed in Table 7.1. However, in Section 7.6.5.1 it has been mentioned that the value of  $\theta$  that is to be used in conjunction with the accurate mobility reduction model of equation (7.69), should be  $0.05V^{-1}$  instead of the value from Table 7.1 that is to be used with the mobility reduction model of equation (7.76). Also, the value of  $v_{sat}$  that will be used in the next sections is  $10^5 m/s$ , which is the value that is assumed to be the saturation velocity in silicon at room temperature. This value will be used in conjunction with velocity-field model 1 of equation (7.84). This is about two times smaller than the fitted value from Table 7.1 that is to be used in conjunction with the velocity-field model 3 of equation (7.86).

#### 7.7.4.1 Nonlinearity coefficients for the triode region

First, the variation of the nonlinearity coefficients is investigated as a function of the channel length. To this purpose, the nonlinearity coefficients that correspond to derivatives with respect to one single voltage  $v_{GS}$ ,  $v_{DS}$  or  $v_{SB}$ , are evaluated for three transistors with a fixed value of W/L but with a different channel length. Also, the bias point is identical for the three transistors. For the three transistor sizes the bias point corresponds to an operating point in the triode region far enough from the saturation region. Figure 7.18 compares the nonlinearity coefficients obtained for a channel length of  $3\mu m$  to the nonlinearity coefficients computed for a channel length of  $1\mu m$  and  $0.7\mu m$ . The values for the  $3\mu m$  transistor have been taken as a reference and the figure shows the relative error with the coefficient values obtained for the two other channel lengths.

Velocity saturation is modeled with the function hot of equation (7.104). This function depends on  $v_{DS}$  only since we neglect the dependence on  $v_{GS}$  and  $v_{SB}$  through the critical field  $E_c$ . Hence, the derivatives with respect to  $v_{GS}$  and  $v_{SB}$  are all given by the function hot multiplied with the derivatives of the product  $large \cdot mobred$ . As a result, the relative error for a given channel length is the same for the derivatives with respect to  $v_{GS}$  only and  $v_{SB}$  only.

Consider now the coefficients  $g_o$ ,  $K_{2g_o}$  and  $K_{3g_o}$ , which are proportional to derivatives of the current with respect to  $v_{DS}$ . It is seen that the error between the coefficients for a  $3\mu m$  channel and shorter channels  $(1\mu m)$  and  $(0.7\mu m)$  increases with the order of the derivative. Recall that with the simple SPICE level 1 model the drain current depends on the square of  $v_{DS}$ . Hence, the nonlinearity coefficient  $K_{3g_o}$  is zero with this model. When the variation of the depletion layer width along the channel and the mobility reduction dependence on  $v_{DS}$  are taken into account, then a nonzero  $K_{3g_o}$  is found. However, it is seen that for short-channel transistors the influence of velocity saturation on  $K_{3g_o}$  is very large: the error between  $K_{3g_o}$  for a  $3\mu m$  transistor and  $K_{3g_o}$  for a  $1\mu m$  transistor is more than 250 %, while for a  $0.7\mu m$  transistor this error increases above 300 %! This indicates that velocity saturation is the main effect that determines  $K_{3g_o}$ .

## 7.7.4.2 Approximate expressions for the nonlinearity coefficients in the triode region

Using the approach explained in Section 3.5 it is possible to find approximate closed-form expressions for the different nonlinearity coefficients. This will be illustrated with the nonlinearity coefficients of the drain current as a function of the voltages  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ . The approximation will be made for a transistor in the  $0.7\mu m$  process with  $L=0.7\mu m$ ,  $W=16\mu m$ ,

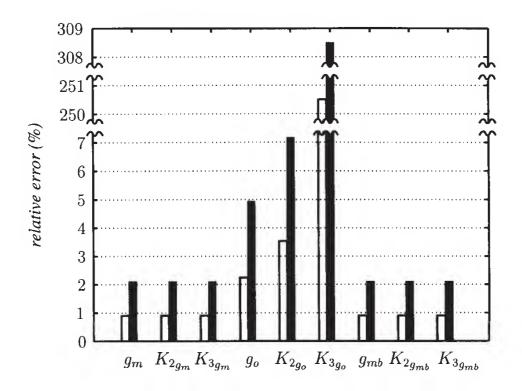


Figure 7.18: Relative error in % between the nonlinearity coefficients of a n-MOS transistor in the triode region (0.7 $\mu$ m technology, W/L=160/7,  $v_{DS}=0.25V$ ,  $v_{GS}=1.9V$ ,  $v_{SB}=1.3V$ ) with  $L=3\mu m$  and the nonlinearity coefficients for a transistor with the same bias and W/L but with  $L=1\mu m$  (white) and  $L=0.7\mu m$  (black). The vertical scale has been broken, in order to represent the low errors and the very high errors on one plot.

 $V_{GB}=3.2V,\,V_{DB}=1.55V$  and  $V_{SB}=1.3V$ . For this transistor it is found that  $V_T=1.17V$  and  $v_{DSAT}=0.55V$ . This value has been obtained by iteration as explained in Section 7.7.3.3.

Table 7.9 lists approximate expressions for the nonlinearity coefficients of the given transistor that is biased in the given bias point of the triode region. Also, the deviation from the exact expression is given. This deviation is expressed as the signed relative error:

signed relative error [%] = 
$$\frac{(approximate\ value) - (exact\ value)}{exact\ value} \cdot 100$$
 (7.122)

approximation is made for a transistor with $L=0.7\mu m$ , $W=16\mu m$ , $V_{GB}=3.2V$ , $V_{DB}=1.55V$ and $V_{SB}=1.3V$ . The $V_{BB}=1.55V$ and $V_{BB}=1.55V$ and $V_{BB}=1.55V$ and $V_{BB}=1.5V$ . The $V_{BB}=1.55V$ and $V_{BB}=1.5V$ . The $V_{BB}=1.5V$ and $V_{BB}=1.5V$ and $V_{BB}=1.5V$ and $V_{BB}=1.5V$ . The $V_{BB}=1.5V$ and $V_{BB$	<b>2</b> <i>Table 7.</i>
$16\mu m$ , $V_{GB}=3.2V$ , $V_{DB}=1.55V$ and $V_{SB}=1.3V$ . The $m$ and $m$ and $m$ and $m$ is $m$ and $m$ and $m$ and $m$ are $m$ and $m$ and $m$ are $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ and $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ and $m$ are $m$ and $m$ are $m$ are $m$ are $m$ and $m$ are $m$ are $m$ are $m$ are $m$ and $m$ are $m$ are $m$ are $m$ are $m$ and $m$ are	Table 7.9: Approximate expressions of nonlinearity coefficients of an n-MOS transistor in strong inversion (triode region).

next page	continued on next page		
		$+rac{1}{6}\cdotrac{\partial^{2}large}{\partial v_{DB}^{2}}\cdotrac{\partial mobred}{\partial v_{DB}}\cdot hot$	<sup>5</sup> 9md
19.1 %	$-15.7 \ \mu A/V^3$	$\frac{\frac{1}{2} \cdot mobred \cdot \left(\frac{1}{3} \cdot \frac{\partial^{3} large}{\partial v_{DB}^{3}} \cdot hot + \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial^{2} hot}{\partial v_{DB}^{2}} + \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot \frac{\partial hot}{\partial v_{DB}}\right)}$	$K_3$
10.5 %	$-1.29 \ mA/V^2$	$-\frac{1}{2}\mu_0 C'_{ox} \frac{W}{L} \left(1 + \frac{\gamma}{2\sqrt{\phi + V_{DB}}}\right) \cdot \frac{1}{1 + \theta f_\mu} \cdot \frac{1}{\sqrt{1 + \left(\frac{V_{DS}}{LE_c}\right)^2}}$	$K_{2g_{md}}$
-3.5 %	0.754~mA/V	$\mu_0 C'_{ox} \frac{W}{L} \left( V_{GB} - V_{FB} - \phi - V_{DB} - \gamma \sqrt{\phi + V_{DB}} \right) \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \frac{1}{\sqrt{1 + \left( \frac{V_{DS}}{LE_c} \right)^2}}$	$g_{md}$
-2.6 %	$0.877  \mu A/V^3$	$\mu_0 C'_{ox} rac{W}{L} \cdot  heta^2 \cdot rac{1}{(1+ heta f_\mu)^3} \cdot rac{1}{\sqrt{1+\left(rac{V_{DS}}{LE_c} ight)^2}}$	$K_{3g_{mg}}$
-2.6 %	$-19.9 \ \mu A/V^2$	$-\mu_0 C'_{ox} rac{W}{L} \cdot  heta \cdot (V_{DB} - V_{SB}) \cdot rac{1}{(1+ heta f_\mu)^2} \cdot rac{1}{\sqrt{1+\left(rac{V_{DS}}{LE_c} ight)^2}}$	$K_{2g_{mg}}$
-2.6 %	$0.452\ mA/V$	$\mu_0 C'_{ox} rac{W}{L} \left(V_{DB} - V_{SB} ight) \cdot rac{1}{1 +  heta f_\mu} \cdot rac{1}{\sqrt{1 + \left(rac{V_{DS}}{LE_c} ight)^2}}$	$g_{mg}$
error	value		coefficient

		continued from next page	next page
nonlinearity	approximate expression at $V_{GB}=3.2V$ , $V_{DB}=1.55V$ , $V_{SB}=1.3V$	exact	relative
coefficient		value	error
gms	$\mu_0 C'_{ox} \frac{W}{L} \left( V_{GB} - V_{FB} - \phi - V_{SB} - \gamma \sqrt{\phi + V_{SB}} \right) \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \frac{1}{\sqrt{1 + \left( \frac{V_{DS}}{LE_c} \right)^2}}$	$1.32\ mA/V$	-2.7 %
$K_{2gms}$	$-\frac{1}{2}\mu_0 C_{ox}' \frac{W}{L} \left( 1 + \frac{\gamma}{2\sqrt{\phi + V_{SB}}} \right) \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \frac{1}{\sqrt{1 + \left(\frac{V_{DS}}{LE_c}\right)^2}}$	$-0.931~mA/V^2$	-25.4 %
$K_{3gms}$	$-rac{1}{2} \cdot rac{\partial large}{\partial v_{SB}} \cdot mobred \cdot rac{1}{L^2 E_c^2 \left(1 + \left(rac{V_{DS}}{LE_c} ight)^2 ight)^2 ight)}$	$-0.40~mA/V^3$	20.8 %
$K_{2g_{mg}\&g_{ms}}$	$-\mu_0 C'_{ox} rac{W}{L} \cdot rac{1}{1+ heta f_{\mu}} \cdot rac{1}{\sqrt{1+\left(rac{V_{DS}}{LE_c} ight)^2}}$	$-1.73\ mA/V^2$	% 6:9-
$K_{32g_{mg}}$ &gms	$\mu_0 C_{ox}^{\prime\prime} rac{W}{L} \cdot  heta \cdot rac{1}{(1+ heta f_{\mu})^2} \cdot rac{1}{\sqrt{1+\left(rac{V_{DS}}{LE_c} ight)^2}}$	$76.1~\mu A/V^3$	-7.4 %
$K_{3}^{}_{g_{mg}\&2g_{ms}}$	$-\frac{3}{2}\mu_{0}C'_{ox}\frac{W}{L}\left(V_{DB}-V_{SB}\right)\cdot\frac{1}{1+\theta f_{\mu}}\cdot\frac{1}{L^{2}E_{c}^{2}\left(1+\left(\frac{V_{DS}}{LE_{c}}\right)^{2}\right)^{3/2}}$	$-0.384 \; mA/V^3$	15.6 %
		continued on next page	next page

							8
$K_{3g_{mg}\&g_{ms}\&g_{md}}$	$K_{3g_{ms}\&2g_{md}}$	$K_{32gms} \& g_{md}$	$K_{2g_{ms}\&g_{md}}$	$K_{3g_{mg}\&2g_{md}}$	$K_{32g_{mg}} k_{g_{md}}$	$K_{2g_{mg}\&g_{md}}$	nonlinearity coefficient
$3 \cdot mobred \cdot \frac{\partial^2 large}{\partial v_{GB} \partial v_{DB}} \cdot \frac{\partial hot}{\partial v_{SB}}$	$\frac{mobred}{2L^{2}E_{c}^{2}\left(1+\left(\frac{V_{DS}}{LE_{c}}\right)^{2}\right)^{3/2}}\left(-6\frac{\partial large}{\partial v_{DB}}\cdot\frac{V_{DS}^{2}}{L^{2}E_{c}^{2}}\left(1+\left(\frac{V_{DS}}{LE_{c}}\right)^{2}\right)}-\frac{\partial large}{\partial v_{SB}}\right)$ $+\frac{\partial^{2} large}{\partial v_{DB}^{2}}\cdot V_{DS}-9 \ large\cdot\frac{V_{DS}}{L^{2}E_{c}^{2}}\left(1+\left(\frac{V_{DS}}{LE_{c}}\right)^{2}\right)$	$rac{\partial large}{\partial v_{SB}} \cdot mobred \cdot rac{1}{L^2 E_c^2 \left(1 + \left(rac{V_{DS}}{LE_c} ight)^2 ight)^{3/2}}$	$\frac{mobred}{L^{2}E_{c}^{2}\left(1+\left(\frac{V_{DS}}{LE_{c}}\right)^{2}\right)^{3/2}}\left(\frac{\partial large}{\partial v_{SB}}\cdot V_{DS}+large\right)$	$rac{\partial^2 large}{\partial v_{GB}\partial v_{DB}} \cdot \left(rac{\partial mobred}{\partial v_{DB}} \cdot hot + rac{\partial hot}{\partial v_{DB}} \cdot mobred ight)$	$-\mu_0 C'_{ox} rac{W}{L}  heta \cdot rac{1}{\left(1 +  heta f_\mu ight)^2} \cdot rac{1}{\sqrt{1 + \left(rac{V_{DS}}{L E_c} ight)^2}}$	$\mu_0 C'_{ox} rac{W}{L} \cdot rac{1}{1+ heta f_\mu} \cdot \sqrt{1+\left(rac{V_{DS}}{LE_c} ight)^2}$	approximate expression at $V_{GB}=3.2V$ , $V_{DB}=1.55V$ , $V_{SB}=1.3V$
$0.614 \ mA/V^3$	$0.408\ mA/V^3 -17.5\ \%$	$-0.797 \ mA/V^3$	$0.353 \ mA/V^2$	$-0.228 \ mA/V^3$	$-78.6 \ \mu A/V^3$	$1.78 \ mA/V^2$	exact value
-5.5 %	-17.5 %	20.5 %	20.0 %	-19.7 %	-4.0 %	-4.5 %	relative error

We now interpret the expressions given in Table 7.9. In addition, we will explain how the approximate expressions of Table 7.9 are obtained with the approach explained in Section 3.5.

Approximate expressions for  $g_{md}$ ,  $K_{2g_{md}}$  and  $K_{3g_{md}}$ . According to equations (7.8) through (7.10), these coefficients are also equal to  $g_o$ ,  $K_{2g_o}$  and  $K_{3g_o}$ .

First we consider the coefficient  $g_{md}$ . This is the first-order derivative of the drain current with respect to  $v_{DB}$ . This can be computed as follows:

$$g_{md} = \frac{\partial i_D}{\partial v_{DB}} = \frac{\partial}{\partial v_{DB}} \left[ large(v_{GB}, v_{DB}, v_{SB}) \cdot mobred(v_{GB}, v_{DB}, v_{SB}) \cdot hot(v_{GB}, v_{DB}, v_{SB}) \right]$$
(7.123)

This derivative can be elaborated further:

$$g_{md} = \frac{\partial large}{\partial v_{DB}} \cdot mobred \cdot hot + \frac{\partial mobred}{\partial v_{DB}} \cdot large \cdot hot + \frac{\partial hot}{\partial v_{DB}} \cdot large \cdot mobred \quad (7.124)$$

The derivatives in this equation can be obtained by differentiating the expressions for large, mobred and hot, which are given in equations (7.90), (7.95) and (7.104), respectively. A numerical evaluation of the three terms in equation (7.124) for the operating point we are considering, shows that the first term in equation (7.124) is much larger than the two other terms. Hence we can approximate  $g_{md}$  by the first term:

$$g_{md} \approx \frac{\partial large}{\partial v_{DB}} \cdot mobred \cdot hot$$
 (7.125)

This approximate expression shows that  $g_{md}$  is mainly determined by the variation of the function large with  $v_{DB}$ . Using the expressions for large, mobred and hot from equations (7.90), (7.95) and (7.104) yields the expression for  $g_{md}$  given in Table 7.9.

Next we consider the second-order coefficient  $K_{2q_{md}}$ :

$$K_{2g_{md}} = \frac{1}{2} \frac{\partial^2 i_D}{\partial v_{DB}^2} = \frac{1}{2} \frac{\partial g_{md}}{\partial v_{DB}}$$
 (7.126)

The right-hand side of this equation can be computed from equation (7.124):

$$K_{2g_{md}} = \frac{1}{2} \cdot \left[ \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot mobred \cdot hot + \frac{\partial^{2} mobred}{\partial v_{DB}^{2}} \cdot large \cdot hot + \frac{\partial^{2} hot}{\partial v_{DB}^{2}} \cdot large \cdot mobred \right.$$

$$+ 2 \cdot \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot hot + 2 \cdot \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial hot}{\partial v_{DB}} \cdot mobred$$

$$+ 2 \cdot large \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot \frac{\partial hot}{\partial v_{DB}} \right]$$

$$(7.127)$$

Expressions for the derivatives in this equation can be obtained by differentiating the expressions for large, mobred and hot twice. An evaluation of the terms in equation (7.127) for the given operating point shows one dominant term, which can be used as an approximation for  $K_{2g_{md}}$ :

$$K_{2g_{md}} \approx \frac{1}{2} \frac{\partial^2 large}{\partial v_{DB}^2} \cdot mobred \cdot hot$$
 (7.128)

This shows that  $K_{2g_{md}}$  is mainly determined by the variation of the function large with respect to  $v_{DB}$ .

Using equations (7.90), (7.95) and (7.104), this expression for  $K_{2g_{md}}$  reduces to the one given in Table 7.9.

Consider now the third-order coefficient  $K_{3g_{md}}$ . This coefficient is given by

$$K_{3g_{md}} = \frac{1}{6} \frac{\partial^3 i_D}{\partial v_{DB}^3} = \frac{1}{3} \frac{\partial K_{2g_{md}}}{\partial v_{DB}}$$

$$(7.129)$$

The right-hand side of this equation can be found by differentiating the right-hand side of equation (7.127) with respect to  $v_{DB}$ . Elaboration of this derivative yields

$$K_{3g_{md}} = \frac{1}{6} \cdot \left[ \frac{\partial^{3} large}{\partial v_{DB}^{3}} \cdot mobred \cdot hot + \frac{\partial^{3} mobred}{\partial v_{DB}^{3}} \cdot large \cdot hot + \frac{\partial^{3} hot}{\partial v_{DB}^{3}} \cdot large \cdot mobred \right.$$

$$+ 3 \cdot \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot hot + 3 \cdot \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot \frac{\partial hot}{\partial v_{DB}} \cdot mobred$$

$$+ 3 \cdot \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial^{2} mobred}{\partial v_{DB}^{2}} \cdot hot + 6 \cdot \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot \frac{\partial hot}{\partial v_{DB}}$$

$$+ 3 \cdot \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial^{2} hot}{\partial v_{DB}^{2}} \cdot mobred + 3 \cdot large \cdot \frac{\partial^{2} mobred}{\partial v_{DB}^{2}} \cdot \frac{\partial hot}{\partial v_{DB}}$$

$$+ 3 \cdot large \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot \frac{\partial^{2} hot}{\partial v_{DB}^{2}}$$

$$(7.130)$$

A numerical evaluation of the terms of  $K_{3g_{md}}$  in the given operating point shows that there are four dominant terms, which can be used as an approximation for  $K_{3g_{md}}$ :

$$K_{3g_{md}} \approx \frac{1}{2} \cdot mobred \cdot \left( \frac{1}{3} \cdot \frac{\partial^{3} large}{\partial v_{DB}^{3}} \cdot hot + \frac{\partial large}{\partial v_{DB}} \cdot \frac{\partial^{2} hot}{\partial v_{DB}^{2}} + \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot \frac{\partial hot}{\partial v_{DB}} \right)$$

$$+ \frac{1}{6} \cdot \frac{\partial^{2} large}{\partial v_{DB}^{2}} \cdot \frac{\partial mobred}{\partial v_{DB}} \cdot hot$$

$$(7.131)$$

These terms are listed in Table 7.9.

With the simple "level 1" model of equation (7.25), the dependence of the drain current on  $v_{DB}$  is quadratic, which implies that  $K_{3g_{md}}$  is zero, as we found in Table 7.4. When the variation of the depletion layer is taken into account, as we did in Table 7.5, then the dependence of  $i_D$  on

 $v_{DB}$  through  $\frac{3}{2}$  powers in addition to the quadratic dependence, yields a small positive value of  $K_{3g_{md}}$ . When both mobility reduction and velocity saturation are taken into account, as we did in this section, we obtain a negative value for  $K_{3g_{md}}$ . Looking at the four terms of the approximate expression of  $K_{3g_{md}}$  we see that  $K_{3g_{md}}$  is determined by several effects. Further it is seen that the four terms are larger in absolute value than  $K_{3g_{md}}$ . This means that the dominant terms partially cancel. Such cancelation of effects leads to a value for the nonlinearity coefficient that is difficult to predict: the effects that have been taken into account in these computations are just an approximation of the real situation and in reality the cancelation of the effects can give a small value that may exhibit a large relative deviation compared to the expression obtained in Table 7.9. It is even possible that the sign of the real coefficient is different from the (negative) sign we found here. Finally, it should be noted that small nonlinearity coefficients that are determined by the compensation of different effects can be determined by minor effects that have not been modeled here.

The reader may have noticed that we did not explicitly write the expression for  $K_{3g_{md}}$  in terms of model parameters and terminal voltages. If we would do so, we would end up with a lengthy expression which is even much more difficult than the expression shown in Table 7.9. The latter expression has been obtained by preventing the routines that compute the approximate expressions to express the derivatives in terms of model parameters and terminal voltages. In this way, an expression is obtained that qualitatively shows the dependence of a nonlinearity coefficient on certain effects.

**Approximate expressions for**  $g_{mg}$ ,  $K_{2g_{mg}}$  and  $K_{3g_{mg}}$ . According to equations (7.5) through (7.7), these coefficients are also equal to  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$ .

We first consider the coefficient  $g_{mg}$ , which is proportional to the first derivative of the drain current with respect to  $v_{GB}$ . When this derivative is computed from the drain current expression of equation (7.105), then it is seen that  $v_{GB}$  occurs inside the curly braces of this expression, as well as in  $f_{\mu}$  and in  $E_c$ . As a result, an exact expression of the derivative is very complicated. An approximate expression is given in Table 7.9.

It is seen that  $g_{mg}$  or  $g_m$  is primarily determined by the dependence of the function large on  $v_{GB}$ . The dependences of the functions mobred and hot on  $v_{GB}$  give a negligible contribution. Indeed, for the given transistor in the given operating point the signed relative error that is made by neglecting these two dependencies, is only -2.6%.

We see that  $g_{mg}$  is proportional to  $V_{DS}$ . This means that  $g_{mg}$  or  $g_m$  is zero for  $V_{DS} = 0V$ . It should be noted that we are making an extrapolation at this moment: the expression in Table 7.9 is an approximate expression that is valid for  $V_{DS} = 0.45V$ , and care should be taken when this expression is used at other bias voltages. However, it has been controlled that the extrapolation to  $V_{DS} = 0V$  is justified here.

Consider now the second-order nonlinearity coefficient  $K_{2g_{mg}}$  which, according to equation (7.6), is also equal to  $K_{2g_m}$ . The coefficient  $K_{2g_{mg}}$  is proportional to the second-order derivative of the drain current with respect to  $v_{GB}$ . In order to compute this derivative, one should take care to not only differentiate the dominant term that determines  $g_{mg}$ , which is the one given in Table 7.9. The derivative of the terms that have been neglected, should be taken

into account as well. This is a general rule that has to be taken into account when interpreting approximate expressions of derivatives that have been computed with the routines described in Section 3.5. In this case, however, the derivatives of the small terms of  $g_{mg}$  are again negligible compared to the derivative of the dominant term of  $g_{mg}$ . This dominant term is given by (see Table 7.9)

$$g_{mg} \approx \mu_0 C'_{ox} \frac{W}{L} \left( V_{DB} - V_{SB} \right) \cdot \frac{1}{1 + \theta f_\mu} \cdot \frac{1}{\sqrt{1 + \left( \frac{V_{DS}}{LE_c} \right)^2}}$$
(7.132)

Two factors in this expression depend on  $v_{GB}$ , namely the factor  $1/(1+\theta f_{\mu})$  and the function hot through the critical field  $E_c$ . Hence, the derivative of  $g_{mg}$  with respect to  $v_{GB}$  is a sum of two terms: one term is determined by the derivative of  $f_{\mu}$  with respect to  $v_{GB}$ . The other term is determined by the derivative of the function hot with respect to  $v_{GB}$ . The latter derivative is not zero, since hot depends on  $v_{GB}$  through  $E_c$ . However, this term turns out to be very small compared to the term that contains the derivative with respect to  $f_{\mu}$ . This means that  $K_{2g_{mg}}$  is primarily determined by the variation of mobility reduction with  $v_{GB}$ . For  $K_{3g_{mg}}$  or, equivalently, for  $K_{3g_{mg}}$ , the same conclusion holds. Note that if  $\theta$  is zero, then the expressions for  $K_{2g_{mg}}$  and  $K_{3g_{mg}}$  of Table 7.9 are zero. This does not mean that the nonlinearity coefficients are really zero. It merely means that the approximate expressions are not valid anymore, since other terms, determined by the dependence of the function hot on  $v_{GB}$  will become dominant.

Approximate expressions for the normalized second- and third-order nonlinearity coefficients can easily be derived from the expressions of  $g_{mg}$ ,  $K_{2g_{mg}}$  and  $K_{3g_{mg}}$  in Table 7.9. They are given by

$$K'_{2g_{mg}} = K'_{2g_m} \approx \frac{-\theta}{1 + \theta f_{\mu}}$$
 (7.133)

$$K'_{3g_{mg}} = K'_{3g_m} \approx \frac{\theta^2}{(1 + \theta f_u)^2}$$
 (7.134)

The accuracy of these expressions is questioned in [Groen 94], especially at large gate-bulk voltages and for non-uniformly doped substrates. This will be further discussed in Section 7.8.4.

Approximate expressions for  $g_{ms}$ ,  $K_{2g_{ms}}$  and  $K_{3g_{ms}}$ . These coefficients are proportional to derivatives with respect to  $v_{SB}$ . The expressions for the functions large, mobred and hot that are given in equations (7.90), (7.95) and (7.104) are symmetric with respect to source and drain. As a result, the exact expressions for the derivatives of  $i_D$  with respect to  $v_{SB}$  can be found from the derivative of the same order with respect to  $v_{DB}$  by interchanging the role of  $v_{DB}$  and  $v_{SB}$ . This is not necessarily true anymore for the approximate expressions since  $v_{DB}$  and  $v_{SB}$  can have a different numerical value. Nevertheless, it is seen that for the operating point used in Table 7.9 this symmetry still holds for  $g_{ms}$  and  $K_{2g_{ms}}$ . For the approximate expression of  $K_{3g_{ms}}$  given in Table 7.9 this is no longer valid. From this approximate expression it is seen that this coefficient

is mainly determined by the variation of the function large with  $v_{SB}$  and by velocity saturation. Indeed, the last factor of the approximate expression of  $K_{3g_{ms}}$  originates from the second-order derivative of the function hot with respect to  $v_{SB}$ . This derivative is found from equation (7.104):

$$\frac{\partial^2 hot}{\partial v_{SB}^2} = -\frac{1}{\left(1 + \left(\frac{v_{DS}}{LE_c}\right)^2\right)^{3/2}} + \frac{3v_{DS}^2}{\left(1 + \left(\frac{v_{DS}}{LE_c}\right)^2\right)^{5/2}} L^4 E_c^4$$
(7.135)

The first term turns out to be much smaller in the given operating point. Hence  $\partial^2 hot/\partial v_{SB}^2$  can be approximated well by the first term of equation (7.135):

$$\frac{\partial^2 hot}{\partial v_{SB}^2} \approx -\frac{1}{\left(1 + \left(\frac{v_{DS}}{LE_c}\right)^2\right)^{3/2}} L^2 E_c^2$$
 (7.136)

This explains how the expression for  $K_{3g_{ms}}$  in Table 7.9 is found.

Approximate expressions for nonlinearity coefficients proportional to cross-derivatives Finally we consider approximate expressions that are proportional to cross-derivatives. It is seen that the approximate expression for  $K_{2g_{mg}\&g_{ms}}$  is identical to the approximate expression for  $K_{2g_{mg}\&g_{md}}$ , apart from the sign. Their exact numerical values are not exactly opposite, which is due to effects that are not modeled in the approximate expressions of Table 7.5. The interpretation of other nonlinearity coefficients is left to the reader.

## 7.7.4.3 Nonlinearity coefficients for the saturation region

In Section 7.6.3 it has been mentioned already that with the use of the mobility reduction model of equation (7.69),  $v_{DSAT}$  and hence the drain current cannot be computed explicitly. Instead, their value must be obtained by iteration. Nevertheless it is possible to obtain closed-form expressions of the nonlinearity coefficients. These expressions can be obtained using the method explained in Appendix F. The closed-form expressions are functions of the terminal voltages and  $v_{DSAT}$ . If in these expressions an approximate formula for  $v_{DSAT}$ , for example the closed-form expression of equation (7.114), is used, then  $v_{DSAT}$  is eliminated from the expressions for the nonlinearity coefficients.

Just as in the previous section, the nonlinearity coefficients are first computed for different channel lengths. Only the nonlinearity coefficients are computed that correspond to derivatives with respect to  $v_{GS}$  only and to  $v_{SB}$  only. No cross-derivatives are considered and no derivatives with respect to  $v_{DS}$  are considered. The latter ones are determined by effects such as channel-length modulation, which are discussed in Section 7.11.

Figure 7.19 shows the relative error in % between the coefficients computed for a channel length of  $3\mu m$  on one hand and the coefficients computed for a channel length of  $1\mu m$  and  $0.7\mu m$ , on the other hand. For the three channel lengths the value of W/L is kept constant and the bias conditions are identical.

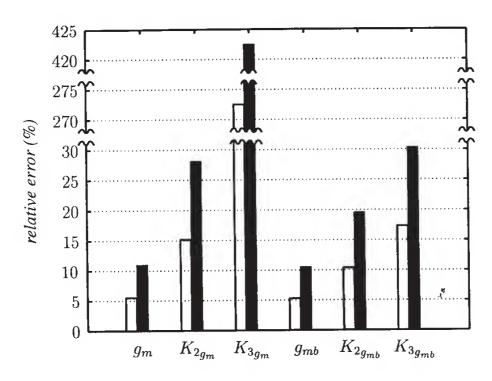


Figure 7.19: Relative error in % between the nonlinearity coefficients of a n-MOS transistor in the saturation region (W/L=160/7,  $v_{DS}=1.45V$ ,  $v_{GS}=1.9V$ ,  $v_{SB}=1.3V$ ) with  $L=3\mu m_e$  and the nonlinearity coefficients for a transistor with the same bias and the same W/L but with  $L=1\mu m$  (white) and  $L=0.7\mu m$  (black). Notice that the vertical scale has been broken, in order to represent the low errors and the very high errors on one plot.

The transistor with a channel length of  $3\mu m$  is long enough such that the influence of velocity saturation, which is more severe for short channels, is reduced. The shorter the transistor is, the larger the influence of velocity saturation is. This is seen in the figure: the deviations on the nonlinearity coefficients increase when the channel gets shorter. Also, the deviations are higher when the order of the nonlinearity coefficient is higher. For example, the error between the coefficient  $K_{3g_m}$  for a  $3\mu m$  transistor and this coefficient for a  $0.7\mu m$  transistor is more than 400%! From this one can conclude that the value of  $K_{3g_m}$  is highly dependent on the velocity saturation effect.

# 7.7.4.4 Approximate expressions for the nonlinearity coefficients in the saturation region

As mentioned above, expressions for the nonlinearity coefficients can be obtained with the approach described in Appendix F. However, these expressions are very complicated. Even if the approximation routines described in Section 3.5 are used, the approximate expressions are still too lengthy to interpret. In order to obtain interpretable expressions we will derive nonlinearity coefficients from the simpler drain current model of equation (7.120) and we will evaluate the error with the coefficients obtained with the approach of Appendix F applied to the more accurate model of equation (7.121). The evaluation will be made for a transistor fabricated in the  $0.7\mu m$  technology of Table 7.1 with  $W=16\mu m$ ,  $L=0.7\mu m$ ,  $V_{GS}=1.9V$ ,  $V_{DS}=1.45V$ 

and  $V_{SB} = 1.3V$ . In this way we find that  $V_T = 1.17V$ ,  $v_{DSAT} = 0.55V$ , and the drain current is 0.385mA. In this value, the effect of the output conductance (channel-length modulation and other effects) is not included. For comparison, the drain current computed with the level 1 model of equation (7.51) is found to be 0.55mA.

**Approximate expression for**  $g_m$  From the model of equation (7.120) we find

$$g_m = \frac{\mu_0 C'_{ox} W}{2a L} \cdot (V_{GS} - V_T) \cdot mobhot^2 \cdot \left(2 + \left(\theta + \frac{\mu_0}{v_{sat} L}\right) \left(V_{GS} - V_T\right)\right)$$
(7.137)

with the function mobhot given by equation (7.119). The value of the function mobhot for the given transistor and the given bias point is 0.763.

The value of  $g_m$  obtained with the model of equation (7.121) is 0.987mA/V for the given transistor in the given bias point. The signed relative error on the value of  $g_m$  obtained with equation (7.137) is -9.4%. This means that the value given in equation (7.137) is a slight underestimation in the given operating point.

**Approximate expression for**  $K_{2g_m}$  Next we evaluate  $K_{2g_m}$  with the drain current model of equation (7.120). We find

$$K_{2g_m} = \mu_0 C'_{ox} \frac{W}{L} \cdot \frac{1}{2a\left(1 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right)^3}$$
(7.138)

With the values from Table 7.1 the signed relative error on equation (7.138) for the given transistor in the given bias point is -23.3% compared to the numerical value obtained with the more accurate drain current model of equation (7.121), which is  $0.523mA/V^2$ . This means that equation (7.138) yields an underestimation of  $K_{2q_m}$  in the given operating point.

The second-order normalized nonlinearity coefficient  $K'_{2g_m}$  is found from equations (7.137) and (7.138):

$$K'_{2g_m} = \frac{1}{\left(1 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right) \cdot \left(2 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right) \cdot (V_{GS} - V_T)}$$
(7.139)

It is seen that mobility reduction and velocity saturation make  $K'_{2g_m}$  smaller. This corresponds to the assertion that a submicron transistor behaves more linearly than a long-channel transistor that satisfies the square-law model [Gar 87b].

For the  $0.7\mu m$  process of Table 7.1 we have  $\theta=0.079V^{-1}$ ,  $\mu_0=0.047m^2/(V.s)$  and  $v_{sat}=1.94\times10^5m/s$ . For a  $0.7\mu m$  transistor the factor  $\mu_0/v_{sat}/L$  is 0.346. This is about 4.4 times higher than  $\theta$ . Hence, velocity saturation plays a larger role in the saturation region than mobility reduction by the normal field. Further, if for the  $0.7\mu m$  transistor  $V_{GS}=1.9V$  and

 $V_{SB}=1.3V$ , then we find  $V_T=1.17V$ , such that  $V_{GS}-V_T=0.73V$ . Then we find that the approximate value of the normalized second-order nonlinearity coefficient  $K'_{2g_m}$  is  $0.45V^{-1}$ . Without mobility reduction and velocity saturation the second-order nonlinearity coefficient would be approximately  $1/2(V_{GS}-V_T)$  (see equation (7.55)), which equals  $0.68V^{-1}$ .

**Approximate expression for**  $K_{3g_m}$  The numerical value of  $K_{3g_m}$  obtained by using the approach of Appendix F on the drain current model of equation (7.121) is  $-0.154mA/V^3$ .

With the drain current model of equation (7.120) the following expression for  $K_{3q_m}$  is found:

$$K_{3g_m} = -\mu_0 C'_{ox} \frac{W}{L} \cdot \frac{\theta + \frac{\mu_0}{v_{sat}L}}{2a\left(1 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right)^4}$$
(7.140)

For the given trans stor in the given bias point the signed relative error of this expression compared to the numerical value obtained with the more accurate drain current model of equation (7.121) is -15.5% which means that using equation (7.140) we obtain a coefficient which is too small in absolute value.

The normalized nonlinearity coefficient  $K'_{3g_m}$  is found by combining equation (7.137) with equation (7.140). This yields

$$K'_{3g_{m}} = -\frac{\theta + \frac{\mu_{0}}{v_{sat}L}}{\left(1 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right)(V_{GS} - V_{T})\right)^{2} \cdot \left(2 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right)(V_{GS} - V_{T})\right) \cdot (V_{GS} - V_{T})}$$
(7.141)

It is seen that this nonlinearity coefficient is completely determined by mobility reduction and velocity saturation in the sense that the coefficient would be zero if these two effects were not present. In that case,  $\theta$  would be zero and  $v_{sat}$  would be infinite. This is not exact: equation (7.141) is only an approximation that neglects the third-order derivative of the function large that determines the drain current for a long-channel transistor with a thick gate oxide. This derivative is not zero, due to the  $\frac{3}{2}$  powers in the function large.

Approximate expression for  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  The numerical values obtained with the model of equation (7.121) for  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  at the given bias point are  $0.249mA/V - 58.6\mu A/V^2$  and  $9.35\mu A/V^3$ , respectively.

Using the simpler drain current model of equation (7.120) combined with the routines described in Section 3.5 to find approximate expressions of derivatives, we find

$$g_{mb} \approx \frac{1}{2a} \mu_0 C'_{ox} \frac{W}{L} \left( v_{GS} - V_T \right) \cdot \frac{\gamma}{\sqrt{\phi + v_{SB}}} \cdot mobhot$$
 (7.142)

The signed relative error on this expression compared to the numerical value obtained with the more accurate drain current model of equation (7.121) is 5.6% which means that equation (7.142) yields an overestimation of  $g_{mb}$ .

For  $K_{2q_{mb}}$  we find the following approximate expression

$$K_{2g_{mb}} \approx \frac{1}{8 a} \mu_0 C'_{ox} \frac{W}{L} \cdot \frac{\gamma \cdot mobhot}{\phi + v_{SB}} \cdot \left( -\gamma - \frac{v_{GS} - V_T}{\sqrt{\phi + v_{SB}}} + 2\gamma (v_{GS} - V_T) \cdot \left( \theta + \frac{\mu_0}{v_{sat} L} \right) \cdot mobhot \right)$$

$$(7.143)$$

The signed relative error on this expression compared to the numerical value obtained with the more accurate drain current model of equation (7.121) is -5.13%.

The first two terms in the sum of the expression for  $K_{2g_{mb}}$  are also present for long-channel transistors. Indeed, when these two terms are combined then, apart from the factor a in the denominator, we obtain the expression of  $K_{2g_{mb}}$  given in Table 3.2. The third term has an opposite sign and it is due to mobility reduction and velocity saturation. The meaning of the difference in sign is that the absolute value of  $K_{2g_{mb}}$  is lowered by mobility reduction and velocity saturation, which can be considered as a linearizing effect.

An approximate expression for  $K_{3q_{mb}}$  is given by

$$K_{3g_{mb}} \approx \frac{1}{16 a} \mu_0 C'_{ox} \frac{W}{L} \cdot \frac{\gamma \ mobhot}{(\phi + v_{SB})^2} \cdot \left( \gamma + \frac{v_{GS} - V_T}{\sqrt{\phi + v_{SB}}} \right)$$

$$- 2\gamma (v_{GS} - V_T) \cdot \left( \theta + \frac{\mu_0}{v_{sat} L} \right) \cdot mobhot$$

$$- \gamma^2 \cdot \left( \theta + \frac{\mu_0}{v_{sat} L} \right) mobhot \sqrt{\phi + v_{SB}}$$

$$(7.144)$$

The first two (positive) terms between the brackets are also present for long-channel transistors with a thick gate oxide, whereas the last two terms, which are negative, are due to mobility reduction and velocity saturation. Again it is seen that these effects decrease the absolute value of the nonlinearity coefficient.

The signed relative error on this expression compared to the numerical value obtained with the more accurate drain current model of equation (7.121) is only 1% in the given operating point. This high accuracy should not imply the too optimistic conclusion that the model equation (7.120) is very accurate. The good accuracy is merely due to a compensation of errors.

**Approximate expression for**  $K_{2g_m\&g_{mb}}$ ,  $K_{32g_m\&g_{mb}}$  and  $K_{3g_m\&2g_{mb}}$  Using the approximation routines explained in Section 3.5 we find for  $K_{2g_m\&g_{mb}}$ 

$$K_{2g_{m}\&g_{mb}} \approx \mu_{0}C'_{ox}\frac{W}{aL} \cdot \frac{\gamma \ mobhot}{\sqrt{\phi + v_{SB}}} \cdot \left(-\frac{1}{2} + (v_{GS} - V_{T}) \cdot \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right) \cdot mobhot\right)$$
(7.145)

Again it is seen that the long-channel value, that corresponds to the factor  $-\frac{1}{2}$  between the brackets, is partially compensated by a contribution of mobility reduction and velocity saturation.

The exact value obtained using the model of equation (7.121) is  $-0.258mA/V^2$ . The signed relative error on expression (7.145) is -28%. This means that the absolute value of  $K_{2g_m\&g_{mb}}$  is underestimated with equation (7.145).

For  $K_{3_{2g_m}\&g_{mb}}$  we find in the same way

$$K_{3_{2g_m \& g_{mb}}} \approx \frac{3}{4 a} \mu_0 C'_{ox} \frac{W}{L} \cdot \left(\theta + \frac{\mu_0}{v_{sat}L}\right) \cdot \frac{\gamma \ mobhot^2}{\sqrt{\phi + v_{SB}}} \cdot \left(1 - 2(v_{GS} - V_T) \cdot \left(\theta + \frac{\mu_0}{v_{sat}L}\right) \cdot mobhot\right)$$
(7.146)

The value of  $K_{3_{2g_m\&g_{mb}}}$  obtained with the drain current model of equation (7.121) is  $0.116mA/V_1^3$ . The signed relative error on expression (7.146) is -21% which indicates an underestimation of  $K_{3_{2g_m\&g_{mb}}}$ .

 $K_{3_{2g_m\&g_{mb}}}$ . When the drain current model of equation (7.120) is used to compute a value for the coefficient  $K_{3_{g_m\&2g_{mb}}}$  then an error of more than 200% is found compared to the value derived from the model of equation (7.121). The value obtained with the latter model is  $-4.63\mu A/V^3$ . Are expression for this nonlinearity coefficient with an acceptable accuracy is very complicated.

# 7.8 Nonuniform doping effects

In modern MOS processes an ion implantation is performed between the source and drain areas in order to adjust the threshold voltage. As a result, the doping concentration can no longer be considered as being uniform in the substrate: it now depends on the depth below the silicon surface. This means that the depletion layer underneath the channel stretches into a region that has a variable doping concentration. As a result, the body-effect coefficient, which depends on the concentration, depends on the width of the depletion layer, and hence on the voltage difference between a point in the channel and the bulk. Then  $\gamma$  becomes dependent on  $v_{DB}$  and  $v_{SB}$ . This qualitative description is now considered in some more detail.

The distribution of the ion concentration after implantation has a shape that resembles to a gaussian profile [Sze 85]. In computations the doping profile is very often approximated by a box profile, as shown in Figure 7.20. The box which stretches from the surface to a depth  $y_I$  is

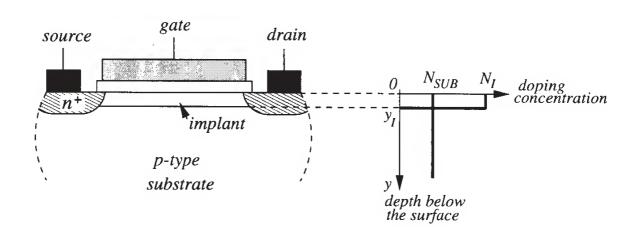


Figure 7.20: A MOS transistor with an ion-implanted channel.

assumed to have a doping concentration  $N_I$ . The net concentration in the region between 0 (the surface) and  $y_I$  is then the sum of this concentration and the concentration  $N_A$  that was present there before the implantation. Deeper than  $y_I$  the concentration is unaffected by the implantation step and so it is also equal to  $N_A$ .

Consider now the operation of the transistor in strong inversion. Below the inversion layer there is a depletion layer. If at any place in the inversion layer the voltage difference with the bulk is very small, then the depletion layer does not stretch beyond  $y_I$ . This situation is depicted in part (a) of Figure 7.21. In this case, the drain current can be computed with the equations presented in the previous sections. Of course, some parameters must now be adapted to a higher doping level of  $N_S = N_A + N_I$ . In this way, new values for  $\gamma$ ,  $\phi$  and  $V_{FB}$  should be used. These

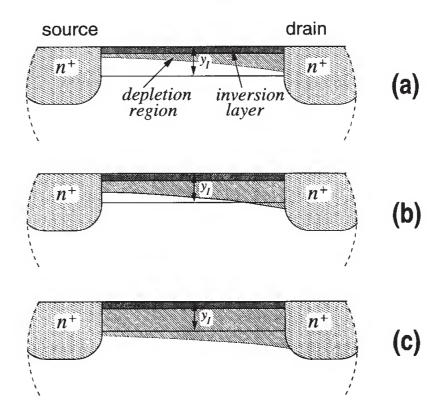


Figure 7.21: Different possibilities for the extension of the depletion layer in a nonuniform substrate: (a) the depletion layer does not extend beyond  $y_I$  anywhere along the channel ( $v_{SB} < v_{DB} < V_I$ ); (b) at the drain side the depletion layer stretches beyond  $y_I$  ( $v_{SB} < v_I < v_{DB}$ ); (c) along the complete channel the depletion layer stretches beyond  $y_I$  ( $V_I < v_{SB} < v_{DB}$ ). Note that the depletion layer around the  $n^+$  regions is not shown for simplicity.

are given by [Tsiv 88]

$$\gamma_1 = \frac{2\varepsilon_{Si}qN_S}{C'_{ox}} \tag{7.147}$$

$$\phi_1 = 2V_t \ln \frac{N_S}{n_i} + 6V_t \approx \phi \tag{7.148}$$

$$V_{FB1} \approx V_{FB} \tag{7.149}$$

Since  $N_S$  is larger than  $N_A$ ,  $\gamma_1$  is larger than  $\gamma$  without an implanted region, and  $\phi_1$  is larger than  $\phi$  without implantation.

This situation can also be described quantitatively. The width of the depletion layer can be written as a function of the voltage difference  $v_{CB}$  between a point in the channel and the bulk:

$$y_d = \sqrt{\frac{2\varepsilon_{Si} \left(\phi_1 + v_{CB}\right)}{qN_S}}, \qquad y_d \le y_I \tag{7.150}$$

If  $v_{CB}$  is increased to a value such that the depletion layer touches the bottom of the simplified implant profile, then this value can be obtained from equation (7.150) by setting  $y_d = y_I$  and then solving for  $v_{CB}$ . The value of  $v_{CB}$  found in this way is denoted as  $V_I$  and is given by [Tsiv 88]

$$V_I = \frac{qN_S y_I^2}{2\varepsilon_{Si}} - \phi_1 \tag{7.151}$$

When  $v_{CB}$  is increased above this value, then the depletion layer will extend beyond  $y_I$ , into the part of the substrate with doping concentration  $N_A$ . Due to the difference in concentration between the substrate part inside and outside the implanted region, the extension of the depletion layer will vary in a different way now. Indeed, the body-effect coefficient that must be used now is  $\gamma_2$ , given by

$$\gamma_2 = \frac{2\varepsilon_{Si}qN_A}{C'_{ox}} \tag{7.152}$$

which is smaller than  $\gamma_1$ .

At positive values of  $v_{DS}$ , the extension of the depletion layer at the drain side is larger than at the source side. Hence, it is possible that the depletion layer at the source side stretches beyond the bottom of the implant whereas the depletion layer at the source side lies completely in the implanted region. This corresponds to part (b) of Figure 7.21. If  $v_{SB}$  is large enough then the depletion layer at the source side can also stretch beyond the bottom of the implant, as shown in part (c) of Figure 7.21.

Consider the following numerical example. For the  $0.5\mu m$  process, from which some parameters are listed in Table 7.1, the doping concentration of the p-substrate is  $4\times10^{16}cm^{-3}$ . The channel is implanted with p-type impurities using an effective dose of  $2\times10^{17}cm^{-3}$ . The depth of the implant  $y_I$  is  $0.135\mu m$ . With these data, we find  $V_I=2.36V$ ,  $\gamma_1=0.768V^{1/2}$  and  $\gamma_2=0.3137V^{1/2}$ . This means that when  $v_{DB}$  and  $v_{SB}$  remain smaller than 2.36V then the uniform doping approximation is valid with a concentration of  $2.4\times10^{17}cm^{-3}$ .

Having in mind that the body-effect coefficient is smaller when the depletion layer stretches beyond the bottom of the implant, it is not difficult to see that the threshold voltage at a point in the channel will change with  $v_{CB}$  as shown with the line with circles in Figure 7.22. This is in fact a piecewise model, consisting of two parts. At low values of  $v_{CB}$ , corresponding to a depletion layer that does not stretch beyond the implanted region, the threshold voltage increases with a large body-effect coefficient  $\gamma_1$ , since the doping concentration in the implanted region is high. As  $v_{CB}$  increases, such that the depletion region extends below the bottom of the implanted region, the increase of  $V_T$  is less drastic, due to a lower body-effect coefficient  $\gamma_2$ . For the sake of comparison to this piecewise model, the threshold voltage is drawn as well for a uniform substrate with the same doping concentration as the implanted region (curve with the "diamonds"), and for the substrate without the implant (curve with the triangles). Clearly, the threshold voltage of the nonuniform substrate first follows the curve for  $\gamma_1$ , after which the curve for  $\gamma_2$  is followed with an offset.

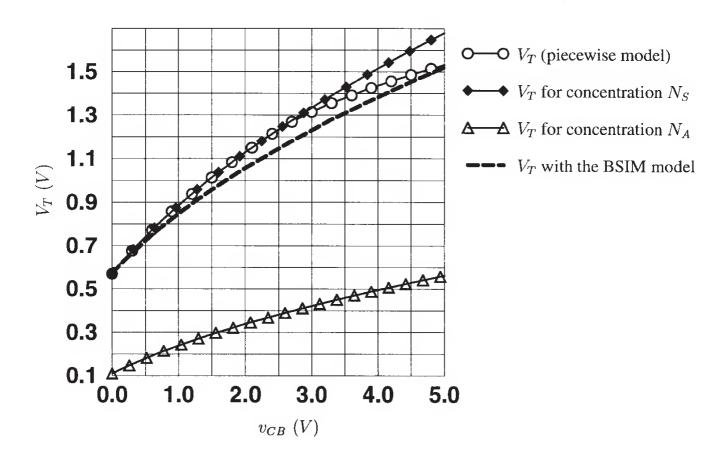


Figure 7.22: Threshold voltage versus reverse bias voltage in the presence of an ion implanted region between source and drain.

## 7.8,1 Modeling with one single body-effect coefficient

In many MOS models the unimplanted-channel model derived in previous sections is used with "compromise" values for the parameters  $\gamma$ ,  $V_{FB}$  and  $\phi$ . For example, the value of  $0.69V^{1/2}$  for  $\gamma$  in Table 7.1 is such a compromise value. This approach can give satisfactory results when the dose of the implantation is low. If, however, derivatives of the drain current are computed in order to obtain the nonlinearity coefficients, then the variation of the body effect coefficient with bias is not taken into account at all. This will yield errors for higher-order derivatives.

## 7.8.2 Adaption of the threshold voltage expression

Most transistor models refer voltages to the source. Also, the threshold voltage is referred to the threshold at the source end of the channel. This can lead to errors on the nonlinearity coefficients as we have seen in previous sections. In such models the nonuniform doping effect is modeled by representing the threshold voltage as follows [Sheu 87, Gow 93, Pow 92, BSIM 95]

$$V_T = V_{TO} + K_1 \cdot \left(\sqrt{\phi + v_{SB}} - \sqrt{\phi}\right) + K_2 \cdot v_{SB}$$
 (7.153)

This is an empirical formula, and the higher-order derivatives of this expression with respect to  $v_{SB}$  will be wrong. For example, the last term in equation (7.153) is linear in  $v_{SB}$ . This is not a correct dependence. As a result of this linear dependence, the influence of this term will disappear after the first derivative of  $V_T$ . This will yield errors on nonlinearity coefficients. The higher the derivative, the larger the errors will be.

For the  $0.5\mu m$  process, the parameters  $K_1$  and  $K_2$  are equal to  $0.637V^{1/2}$  and  $1.47\times 10^{-4}$ , respectively. For this process, the relationship given in equation (7.153) is compared to different threshold voltage models in Figure 7.22. The variable for the horizontal scale should be  $v_{SB}$  instead of  $v_{CB}$ . It is indeed seen that the model equation (7.153) deviates from the piecewise model with an amount up to 80mV. The correspondence between the two models can of course be improved in some region of  $v_{SB}$  values of interest if the fit factors  $K_1$  and  $K_2$  are adapted accordingly.

### 7.8.3 Drain current model with three equations

In [Tsiv 88] a drain current model is presented in which the nonuniform doping effect is modeled more accurately than with an adapted threshold voltage as in equation (7.153). The model of [Tsiv 88] is the extension of the model equation (7.37). It consists of three equations for the current: each of these equations corresponds to a case of Figure 7.21. The current and its first derivative are continuous at the transition from one equation to another. For higher-order derivatives discontinuities occur. The discontinuities arise from the fact that the doping profile is approximated by a box, rather than a more smooth profile.

The result of using this model for the computation of the nonlinearity coefficients is that corrections on the nonlinearity coefficients will have to be made when a large bulk effect is present. In the triode region, corrections will be required for nonlinearity coefficients that comprise derivatives with respect to  $v_{SB}$  and  $v_{DB}$ . In the saturation region corrections will be required for the derivatives with respect to  $v_{GS}$  as well, since  $v_{DSAT}$  will be determined both by the bulk effect and by  $v_{GS}$ . The corrections will be more pronounced for higher-order derivatives.

Effects that have been discussed in previous sections for unimplanted channels are often modeled in the same way for implanted devices. The values of the parameters that describe the effects are fitted accordingly. This can sometimes lead to inaccuracies on the nonlinearity coefficients, as will be discussed in the next section for the modeling of mobility reduction due to the vertical field.

## 7.8.4 Mobility reduction

Mobility reduction in implanted devices is often modeled with the same equations as described in Section 7.6 or with equations that closely resemble [Wu 85] to the ones presented there.

In [Groen 94] measurements of nonlinearity coefficients are presented for a transistor that operates in strong inversion, both in the triode region and in the saturation region. The measurements have been performed on a long-channel device ( $W=10\mu m$  and  $L=20\mu m$ ). In this way, the influence of velocity saturation is minimized. Other short-channel effects that will be discussed in the subsequent sections, are suppressed as well in this way. The measurements

have been performed up to high voltages: for example, the nonlinearity coefficients  $K_{2g_{mg}}$  and  $K_{3g_{mg}}$  have been measured in the triode region for  $V_{GS}$  values up to 14V. In order to obtain a good match with measurements a new model for mobility reduction has been developed. This model takes into account the finite thickness of the inversion layer. The inversion layer stretches in a region with a nonuniform doping concentration: due to the channel implantation, the doping profile is gaussian. Since the mobility of electrons in the inversion is dependent on the doping concentration, the mobility is dependent on the depth beneath the oxide-silicon interface. Since the depth of the inversion layer is dependent on the normal electric field, and therefore dependent on bias, the mobility depends on bias. This reasoning results into a complicated model for the drain current that must be evaluated by numerical integration.

This complicated model has proven to be accurate for long-channel transistors at high bias values. At low bias values simpler models as the ones described in Section 7.6 are still sufficient.

## 7.9 Threshold voltage for short- and narrow-channel devices

The region that must be depleted to build up an inversion layer right below the gate oxide is not limited to the gate area itself. Some of the electric field lines emanating from the gate charges terminate on ionized acceptor atoms on the sides, right next to and below the channel. If the channel width is large, then the part of the depletion region on the sides is relatively small. However, for channel widths of only a few micrometers the side parts become a large percentage of the total. Since the gate must deplete a charge that is larger than predicted by the simple theory, the threshold increases. The increase in threshold voltage is described by the following empirical relationship [Tsiv 88, Anto 88, Lak 94, Hspi 96]:

$$\Delta V_T = \frac{\delta \pi \varepsilon_{Si}}{4C_{ox}'W} \left(\phi + v_{SB}\right) \tag{7.154}$$

in which  $\delta$  is a fit parameter. Other references [BSIM 95] use a slightly different formulation. From this equation it is seen that the threshold voltage now has a term that is linearly dependent on  $v_{SB}$ , whereas the simple theory only contains a term with the square-root of  $v_{SB}$ . This extra term will only affect the first derivative of the threshold voltage. It is seldom important for the computation of nonlinearity coefficients.

The above considerations indicate that the threshold voltage increases as the channel width decreases. On the other hand, as the channel length decreases, then it is found that the threshold voltage decreases. Moreover, it is found that for transistors with short channels the threshold voltage decreases as  $v_{DS}$  increases.

Several models have been proposed to explain the shift of  $V_T$  for short-channel transistors [Lee 73, Toy 79, Pool 84, Kend 86, Liu 93] leading to different names for this effect: charge sharing, barrier lowering or, in order to model the dependence on  $v_{DS}$ , drain-induced barrier lowering (DIBL).

In [Liu 82, Tsiv 88, Anto 88, Hspi 96] the threshold voltage decrease for short-channel transistors is explained as follows: both at the source side of the channel and at the drain side a

part of the depletion layer charge underneath the channel is not controlled by the gate voltage but by the drain voltage and the source voltage, respectively. For a short-channel transistor these two parts become relatively important. These parts do not have to be included in the calculation of  $V_T$ . In other words, the value of  $V_T$  must be lowered. Using this model, the following semi-empirical expression is obtained for the change  $\Delta V_T$  of the threshold voltage [Liu 82, Tsiv 88, Anto 88, Hspi 96]

$$\Delta V_T = -\sigma v_{DS} \quad \text{with} \quad \sigma = \frac{8.14 \times 10^{-22} \eta}{C'_{ox} L^3}$$
 (7.155)

in which  $\eta$  is a dimensionless fit parameter. This expression is used in the SPICE level 2 and 3 models with  $\eta$  represented by the model parameter ETA. For the  $0.7\mu m$  process this parameter equals 0.0052. For a transistor with a channel length of  $0.7\mu m$  this yields a change in  $V_T$  of 6.2mV per Volt  $v_{DS}$ .

According to the model of equation (7.155) the threshold voltage roll-off is inversely proportional to the third power of L. Also, it is seen that the roll-off is inversely proportional to  $C'_{ox}$ . In other words, as  $t_{ox}$  decreases, the threshold voltage roll-off is lower. This can be explained by the fact that with a smaller  $t_{ox}$  the gate is closer to the channel. As a result, the gate is better able to control the depletion layer charge as opposed to releasing this control to the other structures that surround the channel.

An alternative model explains the  $V_T$  roll-off qualitatively as follows [Tsiv 88]. The situation in the channel is affected by field lines that emanate from nearby structures. In the analysis of long-channel devices, the only structures that have been considered were the gate and the substrate, in this context often referred to as the "back gate". However, in short-channel devices the source and drain are so close to all points in the channel that they can affect the situation in the whole channel, just as the gate does. In other words, source and drain play a role comparable to the role of the gate, in addition to their role in long-channel devices. As a result, field lines emanating from all four structures (gate, bulk, source and drain) and terminating on points in the channel must be considered for an accurate description of the device. Bringing the source and drain regions closer to all points in the channel is similar to bringing the gate closer to the channel. The corresponding increase in drain current is modeled by using an effective threshold voltage that is lower than for long-channel devices.

Finally, the concept of barrier lowering explains the shift of the threshold voltage as follows. As L is decreased, more of the region under the inversion layer is depleted for a given gate potential, compared to a long-channel device. The deeper depletion region is accompanied by a larger surface potential, which makes the channel more attractive for electrons. In other words, an increase in surface potential corresponds to a decrease of the potential energy for electrons. Since the potential energy "barrier" to the entrance of electrons into the channel is lowered, the name "barrier lowering" can be used for this effect.

In [Liu 93] a more accurate model than the one from equation (7.155) is given for the reduction of the threshold voltage of deep submicron devices. The model is derived using quasi-two-dimensional analysis. The exact expression for the threshold voltage change is quite complicated and it has a functional form  $Av_{DS} + B\sqrt{v_{DS}}$ . The model agrees well with device simulations

for channel lengths down to  $0.1\mu m$ . For somewhat longer channels, the model equation for  $\Delta V_T$  simplifies to

$$\Delta V_T = -\Theta(L) \left[ 2 \left( V_i - \phi \right) + v_{DS} \right] \tag{7.156}$$

with

$$\Theta(L) = e^{-L/2l} + 2e^{-L/l} \tag{7.157}$$

In this equation,  $V_j$  is the junction potential of the junction between source or drain and the substrate, and l is a *characteristic length*, which has been found empirically to be

$$l \approx 0.1 \left( x_j \, t_{ox} \, x_{dep}^2 \right)^{1/3} \tag{7.158}$$

in which  $x_j$  is the depth of the source and drain junctions and  $x_{dep}$  is the width of the depletion layer underneath the channel, which for simplicity is assumed to be a constant. In this equation l and  $x_j$  are expressed in micrometers and  $t_{ox}$  in Å.

The meaning of this characteristic length can be interpreted as follows: the larger the channel length is compared to l, the smaller the short-channel effect on  $V_T$  will be. Conversely, the smaller l is, the more the short-channel effect on  $V_T$  is postponed to smaller channel lengths. The characteristic length can be made smaller by using a thinner gate oxide, which is in agreement with the conclusion from equation (7.155), although the dependence on  $t_{ox}$  is different. Also, the characteristic length can be lowered by lowering the junction depth  $x_j$  of the source and drain regions or by increasing the channel doping level such that  $x_{dep}$  is lower. The first option can be achieved by using LDD devices while the other option requires a channel implant with a high doping level (see Section 7.8).

It is seen that the dependence of the threshold voltage roll-off on the channel length is also different compared to the dependence used in equation (7.155): the model of equation (7.156) shows an exponential dependence. Further it is seen that the dependence on  $v_{DS}$  is linear in both cases.

Equation (7.156) is the basis for the model of the threshold voltage shift for short-channel transistors in the BSIM3 model.

The threshold voltage roll-off is assumed to be unrelated to saturation: it is present for  $v_{DS}$  values smaller and larger than  $v_{DSAT}$ . Even for voltages above  $v_{DSAT}$ , the channel area is still directly influenced by field lines emanating from the nearby drain.

Since the threshold voltage roll-off is proportional to  $v_{DS}$ , the derivatives of the current with respect to  $v_{DS}$  will be affected by this effect. The threshold voltage roll-off will not significantly influence these derivatives in the triode region, since other effects will be more dominant. However, for transistors in saturation the threshold voltage roll-off will be one of the effects that determine the output conductance. This will be discussed in more detail in Section 7.11.

## 7.10 Source and drain resistances

The channel of a MOS transistor is connected in series with two resistors: the source resistor and the drain resistor. This is shown schematically in Figure 7.23.

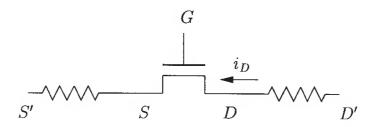


Figure 7.23: A MOS transistor including parasitic source and drain resistances.

As MOS devices scale down, these series resistors become increasingly important. In the equivalent circuit of a bipolar transistor the impedances of the intrinsic device are usually lower (or the conductances) higher than in a MOS transistor. As a result, the relative importance of parasitic resistances in MOS transistors is usually lower than for bipolar transistors. The source and drain resistors consist of different components as will be seen in the next sections.

## 7.10.1 Source and drain resistance components in conventional devices

The source and drain resistances consist of three parts:

- 1. the contact resistance between the  $n^+$  region and metal;
- 2. the resistance of the main body of the  $n^+$  region;
- 3. the resistance associated with the crowding of current flow lines as they go from the  $n^+$  region to the inversion layer, which is usually thinner. This is called the *spreading resistance* effect.

In some transistor models, model parameters are provided to set the source and drain resistances. These resistances are then taken into account by generating additional nodes in the circuit between the intrinsic device and its terminals. However, in several models [Chow 92b, BSIM 95] the current equations are modified to directly include the effect of source and drain resistances. In this way, no additional nodes have to be added, and simulation execution times are reduced [Aro 89]. However, this is not correct for the computation of nonlinearity coefficients as will be discussed in Section 7.10.3.

### 7.10.2 LDD structures

As the channel length of a MOS transistor is reduced, hot-electron effects become more important. Electrons that travel at saturation velocity in the channel can gain enough energy to cause a number of effects. The first effect is *impact ionization*: electron-hole pairs are generated near the drain. The additional electrons contribute to an increased drain current, while the holes cause a substrate current. If the parasitic resistance to the nearest substrate contact is high, the substrate current can lead to a "de-biasing" of the transistor, and possibly snap back [Hsu 83]. The increase of the drain current is further discussed in Section 7.11.

A second effect caused by hot carriers occurs when they have enough energy to surmount the potential barrier at the interface between silicon and silicon oxide. In that case, they can become trapped in the oxide. This results in a cumulative change of device characteristics, which can lead to serious device degradation.

The above effects can be reduced by a reduction of the lateral electric field in the channel. This can be achieved by a modification of the drain structure. Normally, the source is also modified due to the symmetrical nature of MOS transistors, but the main electrical effect is at the drain, where the electric field is usually highest. Such modification can be obtained with the *lightly doped drain (LDD)* structure, that is shown in Figure 7.24. In this structure a lightly-

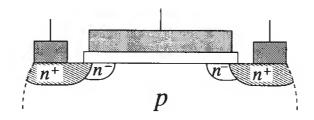


Figure 7.24: A lightly doped drain structure.

doped n-type region is introduced between the channel and the drain. This allows the drain bias to be dropped over a greater distance, and so reduces the drain field. Details about the LDD structure can be found for example in [Hua 87, May 87, Lew 89, Hsu 89].

The lightly-doped drain structure involves an additional series resistance. This degrades the device performance, and the LDD device characteristics need to be optimized to keep the performance penalty minimal. Typically, a 10-20% current and transconductance loss is observed in LDD devices compared to conventional devices [Hsu 89].

The resistance of the LDD regions depends on the transistor terminal voltages. Hence it is nonlinear. A detailed computation of the transistor characteristics that takes into account this nonlinear dependence is found for example in [Hua 87]. However, the analysis is very often simplified in many transistor models as follows: the transistor characteristics are first computed with the assumption of linear resistances [Chow 92b, BSIM 95] but in the final drain current expression the resistance is again allowed to depend on the transistor terminal voltages. This approach yields an acceptable first-order model for the resistance in LDD devices. However, errors will occur in the calculation of nonlinearity coefficients which are proportional to derivatives of the model equations.

## 7.10.3 Effect on the drain current and on the nonlinearity coefficients

The intrinsic drain-source voltage  $v_{DS}$  is smaller than the extrinsic drain-source voltage  $v_{D'S'}$  (see also Figure 7.23):

$$v_{DS} = v_{D'S'} - 2R i_D (7.159)$$

The last term is the voltage drop over the parasitic resistors. In many cases, one can assume that  $Ri_D$  is much smaller than  $v_{GS} - V_T$ . If in addition  $v_{DS}$  is much smaller than  $v_{GS} - V_T$ , then an approximate expression for the drain current in the triode region can be derived from equation (7.25):

$$i_D \approx \frac{W}{L} \mu_{eff} C'_{ox} \left( v_{GS} - V_T \right) v_{DS} \tag{7.160}$$

Combining equations (7.159) and (7.160) yields

$$i_D = \frac{W}{L} \cdot \frac{\mu_{eff} C'_{ox}}{1 + \alpha_R (v_{GS} - V_T)} \cdot (v_{GS} - V_T) v_{D'S'}$$
 (7.161)

with

$$\alpha_R = \frac{2\mu_0 C'_{ox} W R}{L} \tag{7.162}$$

The reduction of the drain current due to the source and drain resistances as described in equation (7.161) has the same form as the simplified model of mobility reduction  $\mu_{eff} = \mu_0/(1 + \theta(v_{GS} - V_T))$ . If both effects are small, then they can be combined as follows:

$$i_D = \frac{W}{L} \cdot \frac{\mu_{eff} C'_{ox}}{1 + (\alpha_R + \theta) (v_{GS} - V_T)} \cdot (v_{GS} - V_T) v_{D'S'}$$
(7.163)

This simplified representation sometimes leads to erroneous interpretations as "mobility reduction due to source and drain resistances", which, of course, make no physical sense.

A computation of the derivatives with the inclusion of the effect of source and drain resistances should better not start from equation (7.163), since it is an oversimplification, both for mobility reduction (see Section 7.6.2) and for the influence of the series resistances. The latter influence is modeled more accurately in references [Chow 92b, BSIM 95] and more specifically for LDD devices in [Hua 87, May 87].

Even if mobility reduction would not be present, which means that  $\theta=0$ , then still equation (7.163) is not a correct starting point for distortion computations. Assume for example that a transistor with a considerable source resistance  $R_S$  is used in a circuit such that a resistor  $R_1$  is placed between its external source and the ground terminal. Then the transistor source degeneration has a loop gain of  $g_m(R_1+R_S)$ . If, one the other hand, equation (7.163) would be used, then the loop gain would be found to be  $g_mR_1/(1+g_mR_S)$ . As a conclusion, it can be stated that for the modeling of source and drain resistances with respect to an accurate computation of nonlinearity coefficients an extra internal source node is best considered, at the expense of more nodes in the complete circuit and, hence, longer CPU times for circuit simulations.

## 7.11 The output conductance and its derivatives in saturation

The output conductance of a MOS transistor is the slope of the  $i_D - v_{DS}$  characteristics with  $v_{GS}$  and  $v_{SB}$  kept constant. When the transistor enters the saturation region then, with the simplified

representation of Figure 7.5, the channel at the drain end gets pinched off, and in first order the drain current becomes independent of  $v_{DS}$ . This would predict a zero slope in the saturation region. In reality the slope is not zero, since there is still a dependence on  $v_{DS}$ . One of the effects that determine this nonzero slope, is the channel-length modulation effect, that has been mentioned already in Section 7.5.1. In the Grove-Frohman model (level 1 from SPICE) this slope is assumed to be constant, as indicated in equation (7.53). Although this model might be sufficient for digital applications, it is far from accurate for analog applications.

The modeling of the output conductance in saturation is not straightforward. It has been investigated by many researchers [Chiu 68, Froh 69, Baum 70, Merck 72, Popa 72, Ross 76, El-Man 77, Tay 79, Poor 80, Tsiv 88, Shou 92, Pow 92, Chow 92b, Hua 92, Enz 95]. One of the reasons for the modeling problems is that at saturation the electric field distribution near the drain end is two-dimensional. The details of the field in that region are affected by the drain region details, such as the junction depth, and also by field lines coming from the gate [Tsiv 88]. An accurate evaluation of the output conductance requires numerical device simulations. This, however, leads to lengthy computer simulations which are not feasible for practical analog integrated circuits. Many efforts have been done to overcome this problem by performing a simplifying pseudo-two-dimensional analysis, that should lead to simplified expressions with an acceptable accuracy. The parameters used in such pseudo-two-dimensional analyses are often empirical or semi-empirical. It is expected that the best models heavily rely on physics rather than on empirical fitting only.

The modeling of the output conductance with the SPICE levels 1, 2 and 3 lack versatility, in these sense that they might be sufficiently accurate only for a small range of transistor dimensions and bias conditions. Moreover, the output conductance according to those models has a discontinuity at the transition from the linear region to saturation. As a result, large errors on the output conductance are reported [Shou 92, Hspi 96], typically 100% or more. It is clear that with such large deviations on the first-order derivative of the drain current with respect to  $v_{DS}$ , the error on the higher-order derivatives can be so high that it is even impossible to just predict an order of magnitude for the value of these higher-order derivatives.

## 7.11.1 The physical model of Huang et al.

In [Hua 92] a physical model for the output conductance is presented that can be applied to both short- and long-channel MOS transistors. The model of Huang is the basis of the modeling of the output conductance in the BSIM3 model. However, in the latter model some fit parameters have been added [BSIM 95].

The modeling of channel-length modulation with the approach of Huang is based on a pseudo-two-dimensional analysis given in [Koh 89]. Similar pseudo-two-dimensional analyses that take into account the effect of lightly-doped drain structures can be found in [May 87, Chow 92a].

The model of Huang does not only take into account the channel-length modulation effect (CLM), it also accounts for drain-induced barrier lowering (DIBL) (see Section 7.9) and the influence of the substrate current (SC). This current is caused by electrons with high energy that travel through the velocity-saturation region: these electrons can impact on atoms and ionize

them, producing electron-hole pairs. This effect is referred to as *impact ionization* and it results in a substrate current. This substrate current flows from the drain to the substrate contact and produces a voltage drop across the substrate resistance. This voltage drop causes a reduction of the body effect, which lowers the threshold voltage and hence increases the drain current [Toh 88, Koh 89]. On the  $i_D - v_{DS}$  transistor characteristics the effect of the substrate current is seen as an increase of the slope at a drain-source voltage typically about 2 V higher than  $v_{DSAT}$ . In order to model this effect, a correction term is added to the drain current:

$$i_{Dtot} = i_D + i_{SUB}$$
 (7.164)

where the first term in the right-hand side is the drain current  $i_D$  without taking into account the impact ionization and the second term models the effect of the substrate current. An expression of  $i_{SUB}$  will be given below.

The derivation of the output conductance with the model of Huang is based upon the drain current model of equation (7.109) which is repeated here for convenience:

$$i_D = \frac{W \mu_{eff} C'_{ox} \left[ (v_{GS} - V_T) v_{DS} - \frac{1}{2} a v_{DS}^2 \right]}{1 + v_{DS} / (LE_c)}$$
(7.165)

and for the saturation region

$$i_{DSAT} = W v_{sat} C'_{or} (v_{GS} - V_T - a v_{DSAT})$$
 (7.166)

with  $v_{DSAT}$  given by

$$v_{DSAT} = \frac{E_c L (v_{GS} - V_T)}{a E_c L + v_{GS} - V_T}$$
 (7.167)

Recall that these drain current expressions have been derived with the piecewise velocity-field model of equation (7.87) and with the assumption that the depletion layer charge varies linearly along the channel.

The derivative of the drain current in saturation with respect to  $v_{DS}$  can be written as the sum of three derivatives of the drain current with respect to  $v_{DS}$ . The three terms are caused by channel-length modulation, drain-induced barrier lowering and by the substrate current:

$$\frac{\partial i_D}{\partial v_{DS}} = \frac{\partial i_D}{\partial v_{DS}} \bigg|_{CLM} + \frac{\partial i_D}{\partial v_{DS}} \bigg|_{DIBL} + \frac{\partial i_D}{\partial v_{DS}} \bigg|_{SC}$$
(7.168)

In accordance with the Early voltage used for bipolar transistors, one can define here an Early voltage for a MOS transistor as well. This Early voltage is defined by

$$V_A = \frac{i_D}{\frac{\partial i_D}{\partial v_{DS}}} \tag{7.169}$$

With each of the terms in equation (7.168) an Early voltage can be associated. Their combination then yields the total Early voltage according to the following equation

$$V_A^{-1} = V_{ACLM}^{-1} + V_{ADIBL}^{-1} + V_{ASC}^{-1} (7.170)$$

in which  $V_{ACLM}$  is the Early voltage caused by channel-length modulation,  $V_{ADIBL}$  is the Early voltage caused by drain-induced barrier lowering and  $V_{ASC}$  is the Early voltage caused by the substrate current. The different contributions to the output conductance or the Early voltage are now discussed.

#### 7.11.1.1 Contribution of channel-length modulation

When  $v_{DS}$  increases above  $v_{DSAT}$  the part of the channel near the drain end in which the electrons travel at maximum velocity, increases towards the source. This reduces the effective channel length which causes an increase in drain current. In [Koh 89] it is shown that the relationship between  $v_{DS}$  and the channel-length reduction  $\Delta L$  is given by

$$v_{DS} = v_{DSAT} + l_t E_c \sinh\left(\Delta L/l_t\right) \tag{7.171}$$

with the parameter  $l_t$  given by

$$l_t = \sqrt{\varepsilon_{Si} t_{ox} x_j / \varepsilon_{ox}} \tag{7.172}$$

and  $x_j$  is the junction depth. For the  $0.5\mu m$  process,  $x_j=0.235\mu m$  and  $t_{ox}=9.4nm$ , such that  $l_t$  equals 81.4nm.

Having a relationship between the channel-length reduction and  $v_{DS}$ , we can compute the contribution of channel-length modulation (CLM) to  $g_o$ . This contribution is given by be

contribution of CLM = 
$$\frac{\partial i_D}{\partial v_{DS}}\Big|_{CLM} = \frac{\partial i_D}{\partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial L} \cdot \frac{\partial L}{\partial v_{DS}}$$
 (7.173)

The three derivatives in the rightmost part of this equation are determined now.

From equation (7.166) we find for the first derivative

$$\frac{\partial i_D}{\partial v_{DSAT}} = -v_{sat}WC'_{ox}a \tag{7.174}$$

The second derivative is found from equation (7.167):

$$\frac{\partial v_{DSAT}}{\partial L} = \frac{E_c \left( v_{GS} - V_T \right)^2}{\left( aE_c L + v_{GS} - V_T \right)^2} \tag{7.175}$$

For the third derivative a simple expression can be obtained when  $\Delta L \gg l$ , which is a valid assumption when  $v_{DS}$  is far enough above  $v_{DSAT}$ . Then equation (7.171) reduces to

$$\frac{v_{DS} - v_{DSAT}}{l_t E_c} = \frac{1}{2} \exp\left(\Delta L/l_t\right) \tag{7.176}$$

It is seen that the channel-length reduction changes with the logarithm of  $v_{DS}-v_{DSAT}$  when the transistor is far enough in saturation. The total effective channel length in saturation is  $L_0-\Delta L$  being the channel length used in the triode region. Hence we find

$$\frac{\partial L}{\partial v_{DS}} = -\frac{\partial \Delta L}{\partial v_{DS}} = \frac{-l_t}{v_{DS} - v_{DSAT}} \tag{7.177}$$

Combining the three derivatives we finally obtain, after some algebra

contribution of CLM = 
$$\left. \frac{\partial i_D}{\partial v_{DS}} \right|_{CLM} = \frac{i_{DSAT} a E_c l_t}{\left( a E_c L + v_{GS} - V_T \right) \left( v_{DS} - v_{DSAT} \right)}$$
 (7.178)

It is seen that this contribution decreases as  $v_{DS}$  increases. At high values of  $v_{DS}$  the contributions from other effects will be more important for short-channel transistors, as will be seen below. For lightly-doped drain structures, the effect of channel-length modulation is smaller [Hua 87, May 87].

In the BSIM3 model the contribution of channel-length modulation, as given in equation (7.178), is multiplied by a fit parameter  $P_{clm}$ . Since the value of  $x_j$  that occurs in the expression of  $l_t$ , cannot be determined very accurately, the parameter  $P_{clm}$  should compensate for the error on the value of  $x_j$ . For the  $0.5\mu m$  process  $P_{clm}$  equals 0.88.

The Early voltage associated with channel-length modulation,  $V_{ACLM}$  is found from equation (7.178) to be

$$V_{ACLM} = \frac{1}{P_{clm}} \frac{aE_cL + v_{GS} - V_T}{aE_cl_t} \left( v_{DS} - v_{DSAT} \right)$$
 (7.179)

### 7.11.1.2 Contribution of drain-induced barrier lowering

In Section 7.9 it has been pointed out that for short-channel devices the threshold voltage decreases about linearly with  $v_{DS}$ . This drain-induced barrier lowering effect (DIBL) also influences  $V_T$  in the saturation regime. Since  $V_T$  changes with  $v_{DS}$ , the drain current changes as well, and in this way DIBL contributes to the output conductance. This contribution is given by

contribution of DIBL = 
$$\frac{\partial i_D}{\partial v_{DS}}\Big|_{DIBL} = \frac{\partial i_{DSAT}}{\partial v_{DS}}$$
 (7.180)

in which  $i_{DSAT}$  is the current in saturation without taking into account the output conductance. Recalling the expression of  $i_{DSAT}$ , equation (7.166), we see that it depends on  $V_T$  not only directly but also through  $v_{DSAT}$ . Hence we obtain

contribution of DIBL = 
$$\left. \frac{\partial i_D}{\partial v_{DS}} \right|_{DIBL} = \frac{\partial i_{DSAT}}{\partial v_{DS}} = \left( \frac{\partial i_{DSAT}}{\partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial V_T} + \frac{\partial i_{DSAT}}{\partial V_T} \right) \frac{\partial V_T}{\partial v_{DS}}$$
 (7.181)

The derivative  $\partial V_T/\partial v_{DS}$  is found using the model equation (7.156) for the drain-induced barrier lowering effect:

$$\frac{\partial V_T}{\partial v_{DS}} = -\left(e^{-L/2l} + 2e^{-L/l}\right) = -\Theta(L) \tag{7.182}$$

in which the function  $\Theta(L)$  is given in equation (7.157). In the BSIM3 model a different function  $\Theta(L)$  is used:

$$\Theta_{BSIM3}(L) = P_{diblc1} \left[ \exp\left(-D_{rout}L/2l_t\right) + 2\exp\left(-D_{rout}L/l_t\right) \right] + P_{diblc2}$$
(7.183)

in which  $P_{diblc1}$ ,  $D_{rout}$  and  $P_{diblc2}$  are fit parameters. For the  $0.5\mu m$  process,  $P_{diblc1}=0.013$ ,  $P_{diblc2}=1.27\times 10^{-3}$  and  $D_{rout}=0.153$ .

The derivative  $\partial i_{DSAT}/\partial v_{DSAT}$  is found from equation (7.174). The derivative  $\partial v_{DSAT}/\partial V_T$  is found from the expression of  $v_{DSAT}$  given in equation (7.167):

$$\frac{\partial v_{DSAT}}{\partial V_T} = -\frac{aE_c^2 L^2}{\left(aE_c L + v_{GS} - V_T\right)^2} \tag{7.184}$$

Combining equations (7.174), (7.181), (7.157) and (7.184) we find

contribution from DIBL = 
$$\left. \frac{\partial i_D}{\partial v_{DS}} \right|_{DIBL} = \frac{i_{DSAT} \left( v_{GS} - V_T + 2aE_cL \right) \Theta(L)}{\left( v_{GS} - V_T \right) \left( aE_cL + v_{GS} - V_T \right)}$$
 (7.185)

and consequently

$$V_{ADIBL} = \frac{(v_{GS} - V_T)(aE_cL + v_{GS} - V_T)}{(v_{GS} - V_T + 2aE_CL)\Theta(L)}$$
(7.186)

The contribution of the DIBL effect to  $g_o$  is seen to be independent of  $v_{DS}$ . This is due to the linear dependence of  $V_T$  on  $v_{DS}$ .

#### 7.11.1.3 Contribution of the substrate current

The substrate current depends on  $v_{DS}$  with the following semi-empirical relationship [Chan 84, Toh 88]

$$i_{SUB} = \frac{A}{B} \cdot i_{DSAT} \cdot (v_{DS} - v_{DSAT}) \cdot \exp\left(-\frac{B \cdot l_t}{v_{DS} - v_{DSAT}}\right)$$
(7.187)

where A and B are empirical constants, and  $l_t$  is given by equation (7.172). It is seen that the dependence of the substrate current on  $v_{DS}$  is almost linear when  $v_{DS}$  is sufficiently higher than  $v_{DSAT}$ .

Taking the derivative of equation (7.187) with respect to  $v_{DS}$  yields the contribution of the substrate current to  $g_o$ :

$$\frac{\partial i_{SUB}}{\partial v_{DS}} = \frac{A}{B} \exp\left(-\frac{B \cdot l_t}{v_{DS} - v_{DSAT}}\right) \cdot i_{DSAT} + \frac{A \cdot l_t}{v_{DS} - v_{DSAT}} \exp\left(-\frac{B \cdot l_t}{v_{DS} - v_{DSAT}}\right) \cdot i_{DSAT} \tag{7.188}$$

The second term is usually negligible for  $v_{DS}$  values far enough above  $v_{DSAT}$ , which is the range where the substrate current becomes significant. Hence we obtain

contribution of 
$$i_{SUB} = \frac{\partial i_{SUB}}{\partial v_{DS}} \approx \frac{A}{B} \cdot i_{DSAT} \cdot \exp\left(-\frac{B \cdot l_t}{v_{DS} - v_{DSAT}}\right) = \frac{i_{SUB}}{v_{DS} - v_{DSAT}}$$
(7.189)

This is an increasing function of  $v_{DS}$ , which is in agreement with the statement that the slope of the  $i_D - v_{DS}$  curve increases at high  $v_{DS}$  values.

The Early voltage associated with the substrate current is given by

$$V_{ASC} = \frac{i_{D_{tot}}}{\frac{\partial i_{SUB}}{\partial v_{DS}}} = \frac{B}{A} \exp\left(\frac{B \cdot l_t}{v_{DS} - v_{DSAT}}\right)$$
(7.190)

In the BSIM3 model the ratio A/B is replaced by  $P_{SCBE2}/L$  and B by  $P_{SCBE1}$ , with  $P_{SCBE1}$  and  $P_{SCBE2}$  being fit parameters. For the  $0.5\mu m$  process of Table 7.1,  $P_{SCBE1} = 4.52 \times 10^8 V/m$  and  $P_{SCBE2} = 5 \times 10^5 m/V$ .

For the BSIM3 model the Early voltage given in equation (7.190) reduces to

$$V_{ASC} = \frac{L}{P_{SCBE2}} \exp\left(\frac{P_{SCBE1} \cdot l_t}{v_{DS} - v_{DSAT}}\right)$$
(7.191)

#### 7.11.1.4 Continuity of the output conductance

One of the pitfalls in many MOS models is the discontinuity of the output conductance at the transition from the triode region to saturation. This problem is solved with the model of Huang by modifying equation (7.170) as

$$V_A = V_{Asat} + \frac{1}{V_{ADIBL}^{-1} + V_{ACLM}^{-1} + V_{ASC}^{-1}}$$
 (7.192)

where  $V_{Asat}$  is the Early voltage at  $v_{DS} = v_{DSAT}$ . This can be obtained from equation (7.165), and it is found to be equal to  $E_cL + v_{DSAT}$ .

Although the continuity of the output conductance has been achieved, this is not the case with the derivatives of the output conductance. This will be shown in the next section with BSIM3 version 2 model parameters. The continuity of the output conductance and its derivatives as computed with the BSIM3 version 3 model has not been investigated yet while this book was written. As a result of the discontinuity, the value of the nonlinearity coefficients such as  $K_{2g_o}$  and  $K_{3g_o}$ , in which higher-order derivatives of the drain current are involved, will still be inaccurate for  $v_{DS}$  values in the vicinity of  $v_{DSAT}$ .

### 7.11.1.5 Evaluation of the output conductance and its derivatives

Having discussed the different contributions to the output conductance,  $g_o$  and the nonlinearity coefficients  $K_{2g_o}$  and  $K_{3g_o}$  can be evaluated. The evaluation has been performed with the BSIM3

version 2 model. The accuracy on the higher-order derivatives is expected to be not very high, due to the different simplifications that have been made in the derivation of the different contributions. Nevertheless, a good indication of the order of magnitude should be obtained for the nonlinearity coefficients.

The three mechanisms that contribute to the output conductance, have a different dependence on  $v_{DS}$ : the contribution of channel-length modulation is proportional to  $1/(v_{DS} - v_{DSAT})$ , the DIBL contribution is constant and the contribution of the substrate current is an increasing function of  $v_{DS} - v_{DSAT}$ . Depending on the model parameters the dominant contribution to the output conductance can be different in different bias regions.

The output conductance and the higher-order derivatives of the drain current with respect to  $v_{DS}$  are evaluated for an n-MOS transistor with a gate width of  $11.4\mu m$  and a gate length of  $0.5\mu m$ . The model parameters are the ones from the  $0.5\mu m$  process from which some model parameters are listed in Table 7.1. The gate-source voltage is kept fixed to 1.4V and the source-bulk voltage is 0V. The drain-source voltage is swept from 0V to 3.3V. With these data  $v_{DSAT}$  is found to be 502mV.

Figure 7.25 shows the output conductance as a function of the drain-source voltage both with and without the effect of the substrate current. At the transition point from the triode region to

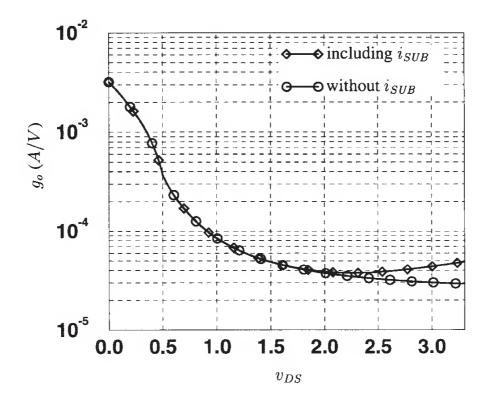


Figure 7.25: The output conductance as a function of  $v_{DS}$ , computed with the BSIM3 version 2 model [BSIM 95] for an n-MOS transistor with  $W=11.4\mu m$ ,  $L=0.5\mu m$ ,  $V_{GS}=1.4V$  and  $V_{SB}=0V$ .

saturation,  $v_{DS} = v_{DSAT}$ , a kink is noticed in the curve of  $g_o$ . As  $v_{DS}$  is increased above  $v_{DSAT}$  it is seen that the output conductance decreases. In this bias region, the output conductance

is primarily determined by channel-length modulation. If the effect of the substrate current is neglected, then it is seen that the output conductance asymptotically goes to a constant value as  $v_{DS}$  increases. This value is determined by the contribution of the drain-induced barrier-lowering effect. However, due to the substrate current the output conductance starts to increase again around  $v_{DS}=2.5V$ .

The nonlinearity coefficient  $K_{2g_o}$  is shown in Figure 7.26 as a function of  $v_{DS}$ . This figure

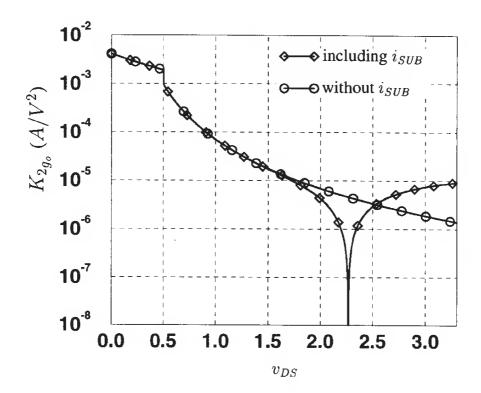


Figure 7.26: Absolute value of  $K_{2g_o}$  as a function of  $v_{DS}$  for the n-MOS transistor with  $W=11.4\mu m, L=0.5\mu m, V_{GS}=1.4V$  and  $V_{SB}=0V$ .

shows a discontinuity at the transition from the triode region to saturation. This has no physical sense and it is due to a shortcoming in the model. At the  $v_{DS}$  value where the output conductance is minimal (see Figure 7.25) we see that  $K_{2g_o}$  is zero. This is clear, since  $K_{2g_o}$  is nothing else but the derivative of  $g_o$  with respect to  $v_{DS}$ , multiplied by two. If the substrate current would be non-existent, then  $g_o$  reaches a constant value at high  $v_{DS}$  values, such that  $K_{2g_o}$  slowly goes to zero.

The third-order nonlinearity coefficient  $K_{3g_o}$  is shown as a function of  $v_{DS}$  in Figure 7.27. Due to the discontinuity of  $K_{2g_o}$  at  $v_{DSAT}$ ,  $K_{3g_o}$  shows a peak at  $v_{DSAT}$ .

It is seen that even with the inclusion of the effect of the substrate current,  $K_{3g_o}$  asymptotically goes to zero for high  $v_{DS}$  values. This is due to the fact that for high values of  $v_{DS}$  the substrate current is almost linearly dependent on  $v_{DS}-v_{DSAT}$ . Due to this linear dependence the derivatives of order higher than one become very small.

Finally, the second- and third-order normalized nonlinearity coefficients of the output conductance are shown in Figure 7.28 and 7.29 as a function of  $v_{DS}$  and for different  $v_{GS}$  values.

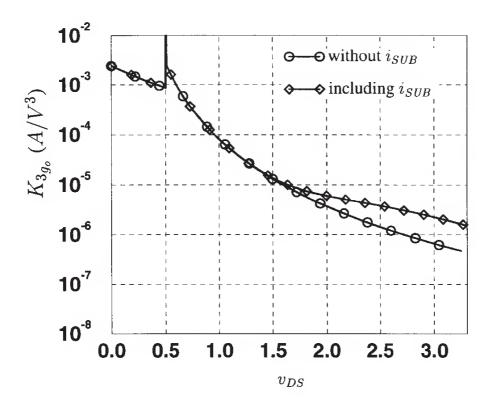


Figure 7.27:  $K_{3g_o}$  as a function of  $v_{DS}$  for an n-MOS transistor with  $W=11.4\mu m, L=0.5\mu m$ ,  $V_{GS}=1.4V$  and  $V_{SB}=0V$ .

The region around  $v_{DSAT}$  should not be considered since the BSIM3 version 2 model is not reliable there. However, it is seen that the normalized nonlinearity coefficients  $K'_{2g_o}$  and  $K'_{3g_o}$  are higher on the average than for a bipolar transistor. Also, the range of variation of these coefficients between  $v_{DSAT}$  and 3.3V is much higher than for a bipolar transistor.

#### 7.11.1.6 Evaluation of other nonlinearity coefficients in saturation

From the model equations that describe channel-length modulation, drain-induced barrier low ering and the substrate current do not only depend on  $v_{DS}$  but also on  $v_{GS}$  and  $v_{SB}$ . Hence the derivatives of the current with respect to  $v_{GS}$  and  $v_{SB}$  are also affected by these effects. However, the dependences of CLM, DIBL and SC on  $v_{GS}$  and  $v_{SB}$  are usually negligible compared for example to the dependences of the functions large, mobred and hot on  $v_{GS}$  and  $v_{SB}$ .

## 7.12 Capacitors in a MOS transistor

In this section the capacitors in a MOS transistor in strong inversion are briefly discussed. Hereby, quasi-static operation of the MOS transistor is assumed. The capacitors that are considered here are shown in Figure 7.1. With these capacitors the quasi-static MOS model is not complete in the sense that the capacitance effect of every terminal of the transistor on every other

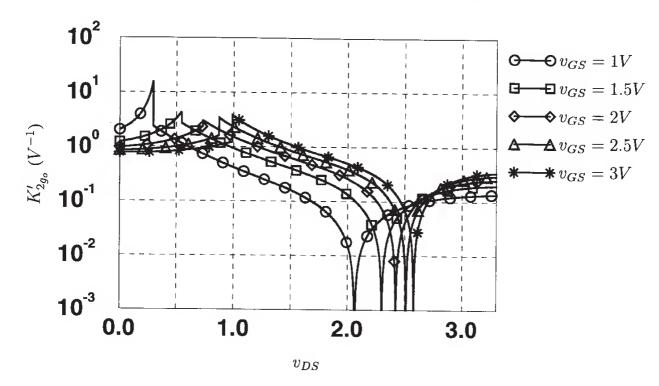


Figure 7.28: Absolute value of  $K'_{2g_o}$  for the n-MOS transistor with  $W=11.4\mu m$  and  $L=0.5\mu m$ , as a function of  $v_{DS}$  and for different values of  $v_{GS}$  ( $v_{SB}=0V$ ).

terminal is not modeled. Such complete modeling requires the use of transcapacitances as explained for example in [Tsiv 88]. However, the capacitances that are shown in Figure 7.1 or in the linearized equivalents of Figure 7.2 and 7.3, are widely used because capacitive effects can be modeled quite simply and with a sufficient accuracy as long as the frequency is not too high.

The capacitors in a MOS transistor can be divided in two classes: extrinsic capacitors and intrinsic capacitors.

# 7.12.1 Extrinsic capacitors

Extrinsic capacitors arise from different sources. First, overlap capacitors arise due to the unavoidable overlap between the gate and the  $n^+$  regions of source and drain and due to the overlap of the gate and the substrate outside the channel region. The latter overlap gives rise to the gate-bulk overlap capacitor  $C_{gbo}$ , while the first overlaps yield the gate-source and the gate-drain overlap capacitors,  $C_{gso}$  and  $C_{gdo}$ , respectively. These capacitors are assumed to be linear.

A second class of extrinsic capacitors consists of the capacitor of the junction between the  $n^+$  source and drain regions and the bulk. These junctions are reversely biased in normal operation. These junction capacitors are of course nonlinear, as described in Section 3.4. These junction capacitors form the first part of  $C_{sb}$  and  $C_{db}$ . The other part of these capacitors originates from the intrinsic part of the transistor, as described in the next section. This part is omitted in many transistor models.

It should be noted that the junction capacitors consist of two parts: a "bottom wall" part and a "sidewall" part. Since the doping concentration of the substrate is higher near the surface, the

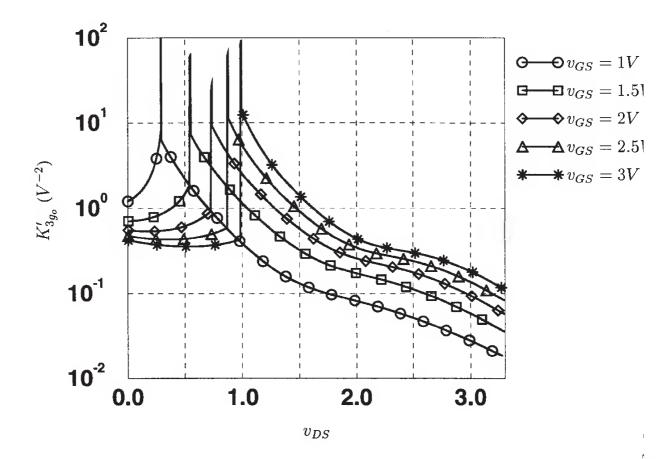


Figure 7.29: Third-order normalized nonlinearity coefficient  $K'_{3g_o}$  for the n-MOS transistor with  $W=11.4\mu m$  and  $L=0.5\mu m$ , as a function of  $v_{DS}$  and for different values of  $v_{GS}$  ( $v_{SB}=0V$ ).

sidewall part is larger per unit area than the bottom wall part. The difference between the two parts gives rise to two different junction potentials and two different grading coefficients.

## 7.12.2 Intrinsic capacitors

The value of the intrinsic capacitors is always a fraction of the total intrinsic capacitance  $C_{ox} = C'_{ox}WL$ . This fraction is bias dependent, and as a result, the intrinsic capacitors are nonlinear. In many MOS models and in hand calculations this bias dependence is neglected. Also, several intrinsic capacitors are completely neglected in many MOS models and hand calculations.

For a MOS transistor with  $W=16\mu m$  and  $L=0.7\mu m$  fabricated in the  $0.7\mu m$  process of Table 7.1,  $C'_{ox}$  is found to be  $2.03fF/(\mu m)^2$ , and the total intrinsic capacitance is 22.75fF. A transistor of minimum channel length and with the same value of W/L but now fabricated in the  $0.5\mu m$  process of Table 7.1 has a width of  $11.4\mu m$ . With  $C'_{ox}=3.45fF/(\mu m)^2$  the total intrinsic capacitance is 19.68fF.

The following intrinsic capacitors can now be distinguished:

1. The gate-source capacitor  $C_{gs}$ . In strong inversion this can be approximated by [Tsiv 88]

$$C_{gs} \approx \frac{2}{3}C_{ox}\frac{1+2\alpha}{(1+\alpha)^2}$$
 (7.193)

with  $\alpha$  given by

$$\alpha = \begin{cases} 1 - \frac{v_{DS}}{v_{DSAT}} & \text{if } v_{DS} < v_{DSAT} \\ 0 & \text{if } v_{DS} \ge v_{DSAT} \end{cases}$$

$$(7.194)$$

From this equation it is found that in the triode region at  $v_{DS}=0V$ ,  $C_{gs}$  equals  $\frac{1}{2}C_{ox}$ , whereas in saturation  $C_{gs}$  equals  $\frac{2}{3}C_{ox}$ . This means that, as  $v_{DS}$  increases,  $C_{gs}$  will increase, until the transistor enters saturation. In other words,  $C_{gs}$  is a capacitance that in the triode depends on the terminal voltages whereas it becomes independent of the terminal voltages as the transistor enters the saturation region. This is not exact. More accurate expressions which, for long-channel transistors, can be found in [Turch 83], indicate that  $C_{gs}$  is not constant in saturation. Also, the dependence of  $C_{gs}$  in the triode region on the terminal voltages is somewhat different than predicted by equation (7.193). For short-channel transistors effects such as velocity saturation influence the capacitance values and should be taken into account [Sheu 84, Iwai 85].

2. The bulk-source capacitor  $C_{sb}^{6}$ . This represents the capacitance of the field-induced junction formed by the substrate and the inversion layer at the source end. It is placed in parallel with the junction capacitor between the  $n^{+}$  source region and the substrate. It is neglected in many models. However, in long-channel transistors it can become significant compared to the junction capacitance between source and drain. Its value is approximately given by [Tsiv 88]

$$C_{sb} = \delta C_{gs} \tag{7.195}$$

where the parameter  $\delta$  has been used in the expression for the drain current, equation (7.44). The parameter  $\delta$  is often given by  $\gamma/\left(2\sqrt{\phi+v_{SB}}\right)$ , as discussed in Section 7.4.3.

At  $v_{DS}=0V$  we find, using equation (7.193), that the capacitance  $C_{sb}$  equals  $\frac{1}{2}\delta C_{ox}$ . Using  $\gamma=\sqrt{2\varepsilon_{Si}N_A}/C'_{ox}$ , we obtain

$$C_{sb}|_{v_{DS}=0V} = \frac{1}{2} \frac{WL\sqrt{q\varepsilon_{Si}N_A}}{\sqrt{2(\phi + v_{SB})}}$$

$$(7.196)$$

This is half the value of the capacitance of an inversely biased junction (see equation (3.107)), where the junction potential is equal to  $\phi$  and the reverse voltage is  $v_{SB}$ . Hence, the shape of the nonlinearity of this capacitance is similar to the shape of the nonlinear junction capacitance between the  $n^+$  source region and the substrate. However, the grading coefficient and the junction potential can be different.

<sup>&</sup>lt;sup>6</sup>In fact we could write the bulk-source capacitor also as  $C_{bs}$ . Since we do not work with transcapacitances here, we do not make a distinction between  $C_{sb}$  and  $C_{bs}$ .

As the transistor enters the saturation region,  $C_{sb}$  is given by

$$C_{sb} = \frac{2}{3}\delta C_{ox} \tag{7.197}$$

Just as for  $C_{gs}$  more accurate expressions for  $C_{sb}$  can be found for example in [Turch 83].

3. The gate-drain capacitor  $C_{gd}$ . Its value is approximately equal to [Tsiv 88]

$$C_{gd} \approx \frac{2}{3} C_{ox} \frac{\alpha^2 + 2\alpha}{\left(1 + \alpha\right)^2} \tag{7.198}$$

with  $\alpha$  given by equation (7.194). Again it is seen that this capacitance depends on the three terminal voltages when the transistor is in the triode region. At  $v_{DS}=0V$  its value equals  $\frac{1}{2}C_{ox}$ . When the transistor is in saturation, its value is zero.

4. The bulk-drain capacitor  $C_{db}$ . This is the capacitance of the field-induced junction formed by the inversion layer at the drain end and the substrate. The value of  $C_{db}$  is approximated by [Tsiv 88]

$$C_{db} = \delta C_{qd} \tag{7.199}$$

Just as with  $C_{sb}$ , this capacitor appears to be parallel with the capacitor between the  $n^+$  region of the drain and the substrate. In saturation we find, using equation (7.198), that  $C_{db}$  is zero.

5. the gate-bulk capacitor  $C_{gb}$ . This capacitor is again dependent on the bias voltages and it is given by

$$C_{gb} = \frac{\delta}{3(1+\delta)} C_{ox} \left( \frac{1-\alpha}{1+\alpha} \right) \tag{7.200}$$

In the triode region this capacitance is zero, while in saturation  $C_{gb}$  equals  $\delta C_{ox}/(3(1+\delta))$ .

The capacitance model of the intrinsic part with the above capacitors is only valid until about one tenth of the frequency  $f_o$  given by [Tsiv 88]

$$f_o = \frac{\mu_{eff}(v_{GS} - V_T)}{2\pi a L^2} \tag{7.201}$$

For a MOS transistor with  $L=0.7\mu m$  fabricated in the given  $0.7\mu m$  technology of Table 7.1 and with  $v_{GS}-V_T=1V$ , and  $v_{SB}=1.3V$ , we find a=1.25 and  $f_o$  is found to be 12.1GHz. Hence the equivalent circuit of Figure 7.1 is valid until about 1.2GHz. At higher frequencies an equivalent circuit must be used that contains transcapacitances or, even better, a non-quasi-static model should be used [Tsiv 88, Park 92, BSIM 95]. These models are not discussed in this book.

Apart from the frequency limitation, the use of the capacitors as in Figure 7.1 has the draw-back that the charge conservation rule is not satisfied. This again can be solved by using transcapacitances [Tsiv 88].

The above analysis of MOS capacitances has shown that in the saturation region the intrinsic capacitors are almost linear. In the calculations in Chapter 8 we will follow this assumption.

# 7.13 Drain current in weak inversion operation

A transistor is said to enter the weak inversion region when  $v_{GS}$  becomes smaller than  $V_T$ . In [Tsiv 88] the upper limit of the weak inversion region for  $v_{GS}$  is given by

$$v_{GS} < V_{FB} + 2\Phi_F + \gamma \sqrt{2\Phi_F + v_{SB}} \tag{7.202}$$

which is smaller than  $V_T = V_{FB} + \phi + \gamma \sqrt{\phi + v_{SB}}$ , since  $\phi$  is larger than  $2\Phi_F$  by a few times  $V_t$ . For the weak inversion region we will consider the drain current and its nonlinearity coefficients.

### 7.13.1 Expression of the drain current

When a potential difference is applied between the gate and the bulk of a MOS transistor then this will cause charges to appear in a region close to the surface. The *surface potential*  $\psi_s$  is defined as the potential drop between the surface and a point outside that region. In strong inversion, when the complete channel is strongly inverted,  $\psi_s$  practically does not change with  $v_{GB}$ , and we have for a point x between 0 and L:

$$\psi_s(x) = \phi + v_{CB}(x) \tag{7.203}$$

It is seen that the potential depends on the place in the channel.

As the gate-source voltage drops below  $V_T$ , then the MOS transistor is said to enter the subthreshold region and the drain current begins to fall rapidly with decreasing  $v_{GS}$ . In this operating region, referred to as the weak inversion operating region, the surface potential now depends on  $v_{GB}$  and is given by [Tsiv 88]

$$\psi_s \approx \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + v_{GB} - V_{FB}}\right)^2 \tag{7.204}$$

It is seen that  $\psi_s$  is now independent of the position in the channel. In other words, the potential difference between two points at the surface is zero. As a result, a nonzero current can only be caused by diffusion of electrons. This contrasts with the transport mechanism of electrons in the strong inversion operation, which is caused by drift, due to a nonzero electric field.

Based on equation (7.204), one can compute an approximate closed-form expression for the subthreshold diffusion current for large-channel devices, which yields [Aro 89, Tsiv 88, Dun 91]:

$$i_D = \frac{W}{L} \frac{\mu_0 C'_{ox} \gamma V_t^2}{2\sqrt{\psi_s}} \exp\left[ (\psi_s - 2\Phi_F)/V_t \right] (1 - \exp(-v_{DS}/V_t))$$
 (7.205)

where  $\psi_s$  is computed using equation (7.204).

The dependence of  $v_{GB}$  and hence  $v_{GS}$  on  $\psi_s$  (equation (7.204)) is almost linear. As a result, equation (7.205) is simplified in many CAD models to an expression of the form [Aro 89, BSIM 95]

$$i_D = I_{s0} \left( 1 - \exp(-v_{DS}/V_t) \right) \exp\left( \frac{v_{GS} - V_T - V_{off}}{nV_t} \right)$$
 (7.206)

Here  $V_{off}$  is a model parameter. For the  $0.5\mu m$  process its value is -0.125V. Further, the parameter n is the so-called *weak-inversion slope* which is the slope of the functional relationship between  $v_{GB}$  and  $\psi_s$ :

$$n = \frac{dv_{GB}}{d\psi_s} \tag{7.207}$$

This is approximately given by by [Tsiv 88]

$$n \approx 1 + \frac{\gamma}{2\sqrt{1.5\Phi_F + v_{SB}}} \tag{7.208}$$

For short-channel devices, and particularly at large values of  $v_{DS}$ , the surface potential  $\psi_s$  is no longer constant along the channel and it is higher than for long-channel devices. Still the exponential relationship applies, but the parameters n,  $I_{s0}$  and  $V_{off}$  are fitted to an acceptable value. This way of working, of course, may give large errors on the derivatives.

The influence of the ion implantation for the  $V_T$  adjust (see Section 7.8) on the transistor characteristics in weak inversion can be described as follows: for large  $v_{SB}$  values, the depletion region is outside the implant and the device qualitatively behaves as an unimplanted device. The slope of the curve  $\log(i_D)$  versus  $v_{GS}$  is proportional to n as before, where n is again given by equation (7.208), but with  $\gamma = \gamma_2$  with  $\gamma_2$  given in equation (7.152). For low values of  $v_{SB}$  the depletion region lies completely inside the implant. Then  $\gamma = \gamma_1$  (equation (7.147)), such that n is larger and the slope is smaller. As  $v_{GS}$  increases, the depletion region can move over a region with a widely varying concentration. Then n becomes dependent on  $v_{GS}$ .

We now compute the weak inversion slope for the  $0.5\mu m$  technology of Table 7.1. It is assumed that the source-bulk voltage difference is small enough such that the depletion layer does not stretch beyond the implantation region for the adjustment of the threshold voltage. In this region, we found in Section 7.8 that the doping concentration is  $2.4 \times 10^{17} cm^{-3}$ . Using equation (7.21) the Fermi potential  $\Phi_F$  is found to be 0.429V. Further,  $\gamma = 0.768V^{1/2}$  as computed in Section 7.8. Then for  $v_{SB} = 1.3V$  we find, using equation (7.208), n = 1.275.

# 7.13.2 Nonlinearity coefficients of the drain current in weak inversion

In order to have an idea about the value of the nonlinearity coefficients that describe the drain current in weak inversion, some derivatives of equation (7.206) are computed.

For the transconductance  $g_m$  we easily find

$$g_m = \frac{i_D}{n V_t} \tag{7.209}$$

and for the second- and third-order nonlinearity coefficients

$$K_{2g_m} = \frac{i_D}{2n^2 V_t^2} \tag{7.210}$$

$$K_{3g_m} = \frac{i_D}{6n^3 V_t^3} \tag{7.211}$$

and for the normalized nonlinearity coefficients

$$K_{2g_m}' = \frac{1}{2n\,V_t} \tag{7.212}$$

$$K_{3g_m}' = \frac{1}{6n^2V_t^2} \tag{7.213}$$

The expressions of the nonlinearity coefficients are closely related to the expressions of the nonlinearity coefficients that describe the dependence of the collector current of a bipolar transistor on  $v_{BE}$ : in both cases we find an exponential dependence of the current on the controlling voltage. An important difference, however, is that the slope of the curve in weak inversion is proportional to 1/n, which is considerably smaller than one, whereas for a bipolar transistor it is proportional to  $1/n_F$  (see equation (6.2)), which is very close to unity under normal injection conditions.

**Derivatives with respect to**  $v_{SB}$  For the computation of the derivatives with respect to  $v_{SB}$ , we start from the current expression of equation (7.205). If  $v_{DS}$  is sufficiently larger than zero, then the exponential term that contains  $v_{DS}$  in equation (7.205) is about zero, and we find

$$g_{mb} \approx i_D \left( -\frac{1}{nV_t} + \frac{1}{V_t} + \frac{1}{2n\psi_s} \right) \tag{7.214}$$

The last term between the brackets is negligible compared to the first two terms, and we find

$$g_{mb} \approx \frac{n-1}{n} \frac{i_D}{V_t} \tag{7.215}$$

An approximate expression for the second-order nonlinearity coefficient is found to be

$$K_{2g_{mb}} \approx \frac{i_D}{V_t^2} \cdot \frac{(n-1)^2}{2n^2}$$
 (7.216)

and for the third-order nonlinearity coefficient

$$K_{3g_{mb}} \approx \frac{i_D}{V_t^3} \cdot \frac{(n-1)^3}{6n^3}$$
 (7.217)

In this way, we find for the normalized nonlinearity coefficients

$$K_{2g_{mb}}' \approx \frac{1}{V_t} \cdot \frac{n-1}{2n} \tag{7.218}$$

$$K'_{3g_{mb}} \approx \frac{1}{V_t^2} \cdot \frac{(n-1)^2}{6n^2}$$
 (7.219)

**Derivatives with respect to**  $v_{DS}$  From equation (7.205) or (7.206) we find for  $g_o$ 

$$g_o = \frac{i_D}{V_t} \cdot \frac{\exp(-v_{DS}/V_t)}{1 - \exp(-v_{DS}/V_t)}$$
(7.220)

This equation predicts that  $g_o$  rapidly goes to zero as  $v_{DS}$  increases. However, equation (7.220) neglects the direct influence of the drain field on the channel, as discussed in Section 7.11. Hence, the use of equation (7.220) to derive the nonlinearity coefficients that are proportional to derivatives of the drain current with respect to  $v_{DS}$ , is questionable.

## 7.14 Summary

In this chapter we discussed the nonlinearity coefficients that describe the different basic nonlinearities of a MOS transistor. Most attention has been paid to the modelling of the drain current in strong inversion. Both the triode region and the saturation region have been considered. Since the drain current is a function of three terminal voltages, nineteen nonlinearity coefficients of order one to three are required to describe the drain current's nonlinearities. First the nonlinearity coefficients have been derived for a long-channel transistor without any second-order effects. In this case, the quadratic model is often used in saturation, although a more accurate model takes into account extra terms with  $\frac{3}{2}$  powers. It is seen that there are already significant differences between the higher-order nonlinearity coefficients for these two models. The quadratic model corresponds to the SPICE level 1 model, while the model with the  $\frac{3}{2}$  powers is a part of the SPICE level 2 model. The level 3 model and the BSIM model approximate the  $\frac{3}{2}$  models by low-order polynomials for efficiency reasons. However, this leads to large deviations for some higher-order nonlinearity coefficients.

Next, different models for mobility reduction have been considered. Again, significant differences have been noticed between the higher-order nonlinearity coefficients obtained with more complicated models and with simplified models.

For velocity saturation the most widely used models have been considered. The nonlinearity coefficients have been computed with a velocity-field model that is more accurate than the piecewise velocity-field model that is widely used in MOS models nowadays. As a result of using this more accurate model, the saturation voltage cannot be computed accurately anymore. Instead it must be computed by iteration. Clearly, such model is not efficient for numerical simulations of large transistor circuits. However, it yields a better accuracy for the nonlinearity coefficients.

Although the drain current model is very complicated when an accurate mobility reduction model and velocity saturation model are taken into account, approximate closed-form expressions have been derived for the nonlinearity coefficients. This has been achieved by the approach described in Section 3.5 that allows to identify the dominant terms of a nonlinearity coefficient. The approximations have been obtained for transistors of a given size in a given bias point. One should be careful when the reported approximate expressions are used for transistors of other technologies than the ones used in this chapter (technologies down to  $0.5 \mu m$ ) and for other bias points. For example, some nonlinearity coefficients are determined by several effects that partially cancel. The net value can then be very small. As a result, the real value of such nonlinearity

7.14 Summary 301

coefficient can largely deviate from the computed value. In addition, the canceling effects may be masked by other effects that have not been taken into account during the calculations.

The accurate mobility reduction model and the velocity saturation model that have been used in this chapter to derive nonlinearity coefficients are symmetrical with respect to source and drain in the triode region of the transistor. This corresponds to the physical reality: at  $v_{DS}=0V$  the role of source and drain is identical. However, such symmetry is seldom used in most MOS models for efficiency reasons. As a result, many MOS models yield inaccuracies on the nonlinearity coefficients in the triode region, especially in the vicinity of  $v_{DS}=0V$ . In the saturation region, however, the effect of this asymmetry is small, since the role of source and drain in this operating region is completely different.

Next, some other effects have been discussed: nonuniform doping effects, the influence of source and drain resistors, the variation of the threshold voltage with bias conditions for short and narrow transistors and the nonlinearity of the output conductance in saturation. This output conductance has been modeled with the model of Huang et al. . This output conductance is caused by several effects: channel-length modulation, drain-induced barrier lowering and the substrate current. By taking into account these effects, the output conductance can be described quite accurately and a realistic order of magnitude can be obtained for the coefficients  $K_{2g_o}$  and  $K_{2g_o}$ .

Finally, it should be noticed that the modeling of a MOS transistor is a difficult, never ending task. Due to the rapid scaling of MOS technologies, additional effects need to be taken into account while effects that have been well modeled for transistors of older technologies need to be modeled in another way for newer technologies. This will of course influence the computations of the nonlinearity coefficients.

## **Chapter 8**

# Weakly nonlinear behavior of basic analog building blocks

### 8.1 Introduction

In this chapter we study the nonlinear distortion in some basic analog building blocks both in MOS and bipolar. Distortion in analog circuits that consist of analog building blocks, has been studied already for some particular classes of circuits, such as time-continuous filters [Groen 94, Tsiv 93a, Tsiv 94]. A treatment of these circuits would require a thorough study of those classes of circuits, which is beyond the scope of this book. Instead, we will analyze some small circuits in order to provide the reader some feeling about nonlinear distortion in general. This experience can be helpful in the analysis or design of more complicated circuits such as filters.

An amplifier with a single bipolar transistor is studied first. This circuit is studied in depth and the calculations are elaborated: in this way, the reader can become familiar with the calculation method explained in Chapter 5 with a small example circuit. In addition, an analysis in depth of this small example circuit will yield insights that are useful for larger circuits as well.

Although the circuit is very small compared to analog integrated circuits of nowadays, the expressions already become very lengthy. Nevertheless, approximate interpretable expressions can be obtained using the approximation facilities of ISAAC. Using ISAAC in this way, we can analyze distortion deeper than in other textbooks or papers.

However, it is not always necessary to have a symbolic expression in order to get insight in distortion mechanisms. In Chapter 5 it has been explained that a given harmonic or intermodulation product can be considered as the sum of contributions of the different nonlinearities in the circuit. A plot of the different contributions, for example as a function of the fundamental frequency, can already be very instructive, such that it is not necessary anymore to derive an expression.

Next, a single-MOS-transistor amplifier will be studied. The calculations will not be made as much in detail as for the bipolar case, since many results can be adopted from the bipolar case. For the evaluation of the nonlinearity coefficients in this example, we will make use of the transistor model that takes into account velocity saturation and mobility reduction, as described

in Sections 7.7.2.4 and 7.7.3.3.

Next, some other basic building blocks are studied: in Sections 8.4 and 8.5 differential pairs in with bipolar and MOS transistors are discussed. Differential pairs ideally do not produce even-order distortion. This is no longer true when mismatches are taken into account. Mismatches are important as well in current mirrors, which will be discussed in Section 8.10.

Further, in Section 8.6 an emitter follower is considered, and a source follower is studied in Section 8.7. The distortion of a cascode transistor driven by a current is studied in Section 8.8. Distortion in a common-gate and a common-base transistor driven by a voltage source is analyzed in Section 8.9. Finally, the nonlinear behavior of current mirrors is studied in Section 8.10.

The capabilities of a symbolic network analysis tool such as ISAAC to obtain insight in the nonlinear distortion of somewhat larger circuits, are illustrated with two examples, namely a bipolar double-balanced mixer based on a Gilbert cell, and a Miller-compensated operational amplifier.

Further, we will study a CMOS upconvertor with a mixer transistor biased in the triode region. Using the model of Section 7.7.2.4 in conjunction with the routines that can derive approximate expressions for nonlinearity coefficients (see Section 3.5), we will analyze the nonlinear behavior of the mixer transistor up to the fourth order. The results will be compared to measurements.

## 8.2 Single bipolar transistor amplifier

Figure 8.1 depicts a single-transistor amplifier that is excited by a voltage source. The output resistance of the voltage source is  $R_S$ , which is assumed to be small. The circuit is first analyzed

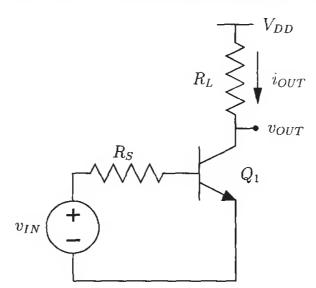


Figure 8.1: A single transistor amplifier (BJT version) with a voltage source excitation.

with  $R_S$  equal to zero and with a transistor model that neglects the parasitic capacitances and ohmic resistances as well as the Early resistance. Afterwards, the transistor model is extended and the effect of a nonzero  $R_S$  is examined. The nonlinearity coefficients of the transistor are

computed with the Gummel-Poon model. The SPICE model parameters used to compute these coefficients are given in Table 8.1.

			2 .
$I_S$	$6.5 \times 10^{-18} A$	$I_{KF}$	$8.0 \times 10^{-3} A$
$eta_F$	120	$I_{RB}$	$1.0 \times 10^{-5} A$
$n_F$	1.0	$r_{BM}$	$120\Omega$
$\beta_R$	10	$r_B$	$378\Omega$
$I_{SE}$	$3.4 \times 10^{-16} A$	$n_E$	2.0
$n_R$	1.0	$C_{JE}$	$2.2 \times 10^{-14} F$
$V_{JE}$	0.9V	$m_{JE}$	0.4
$C_{JC}$	$2.8 \times 10^{-14} F$	$V_{JC}$	0.7V
$m_{JC}$	0.33	$C_{JS}$	$5.0 \times 10^{-14} F$
$V_{JS}$	0.6V	$m_{JS}$	0.33
FC	0.9	$ au_F$	10ps
$I_{TF}$	20mA	XTF	10
$V_{TF}$	1.4V	$V_{AF}$	30V

Table 8.1: SPICE Gummel-Poon model parameters of the bipolar transistor used in this section

## 8.2.1 Elementary transistor model

It is first assumed that the collector current  $i_C$  is only a function of the base-emitter voltage, a described by the following equation (see also Chapter 3):

$$i_C = I_S \exp(\frac{v_{BE}}{V_t}) \tag{8}$$

The output of interest is either the voltage over the (linear) resistor  $R_L$  or the current through  $R_L$ . Since the simple transistor model that is used here does not contain an Early resistance, the transistor behaves like an ideal controlled current source the current of which completely flow through  $R_L$ . Hence, the value of the harmonics and intermodulation products is not affected to  $R_L$  such that an analysis of the harmonic distortion on the AC voltage over  $R_L$  yields the same results as for the harmonic distortion of the AC current through  $R_L$ .

## 8.2.1.1 Computation of harmonics from the DC transfer characteristic

In this simple case it is possible to derive an analytic expression for the input-output relationsh without having to make the assumption of weakly nonlinear behavior. This relationship, of referred to as a DC transfer characteristic, can then be developed into a power series. From coefficients of this series it is possible to derive an expression for the harmonic and intermodution distortion figures.

One easily finds for the DC transfer characteristic

$$v_{OUT} = V_{DD} - R_L I_S \exp(\frac{v_{IN}}{V_t})$$
(8.2)

The input and output voltages are split into a quiescent part and a time-varying part:

$$v_{IN} = V_{IN} + v_{in} \tag{8.3}$$

and

$$v_{OUT} = V_{OUT} + v_{out} (8.4)$$

In this way, one obtains:

$$v_{OUT} = V_{DD} - R_L I_S \exp\left(\frac{V_{IN} + v_{in}}{V_t}\right)$$
(8.5)

$$= V_{DD} - R_L I_C \exp(\frac{v_{in}}{V_t}) = V_{DD} - V_{OUT} \exp(\frac{v_{in}}{V_t})$$
(8.6)

This relationship can be expanded in a Taylor series:

$$v_{OUT} = V_{DD} - R_L I_C \left( 1 + \frac{v_{in}}{V_t} + \frac{1}{2!} \left( \frac{v_{in}}{V_t} \right)^2 + \frac{1}{3!} \left( \frac{v_{in}}{V_t} \right)^3 + \dots \right)$$
(8.7)

From this equation, the AC part can be isolated:

$$v_{out} = -R_L I_C \left( \frac{v_{in}}{V_t} + \frac{1}{2!} \left( \frac{v_{in}}{V_t} \right)^2 + \frac{1}{3!} \left( \frac{v_{in}}{V_t} \right)^3 + \dots \right)$$
(8.8)

This AC part can be identified with the general Taylor series (see also equation (2.4)):

$$v_{out} = K_1 \cdot v_{out} + K_2 \cdot v_{out}^2 + K_3 \cdot v_{out}^3 + \dots$$
 (8.9)

which yields

$$K_1 = -\frac{R_L I_C}{V_t} = -g_m R_L (8.10)$$

$$K_2 = -\frac{R_L I_C}{2V_t^2} = -\frac{g_m R_L}{2V_t} \tag{8.11}$$

$$K_3 = -\frac{R_L I_C}{6V_t^3} = -\frac{g_m R_L}{6V_t^2} \tag{8.12}$$

When the excitation consists of a quiescent part and a sine wave

$$v_{IN} = V_{IN} + V_{in}\cos(\omega_1 t) \tag{8.13}$$

the different harmonics at the output are given by

$$V_{out,1,0} = -g_m R_L V_{in} I_{out,1,0} = -g_m V_{in} (8.14)$$

$$V_{out,2,0} = -\frac{g_m R_L}{4V_t} V_{in}^2 \qquad I_{out,2,0} = -\frac{g_m}{4V_t} V_{in}^2$$
(8.15)

$$V_{out,3,0} = -\frac{g_m R_L}{24V_t^2} V_{in}^3 \qquad I_{out,3,0} = -\frac{g_m}{24V_t^2} V_{in}^3$$
 (8.16)

and the corresponding harmonic distortion figures for both the output voltage and output current are found to be

$$HD_2 = \frac{V_{in}}{4V_t} \tag{8.17}$$

$$HD_3 = \frac{V_{in}^2}{24V_t^2} \tag{8.18}$$

The computations can be extended with the inclusion of a forward emission coefficient. Doing so, one finds

$$HD_2 = \frac{V_{in}}{4n_F V_t} {(8.19)}$$

$$HD_3 = \frac{V_{in}^2}{24n_F^2 V_t^2} \tag{8.20}$$

Under high-injection conditions the collector current is proportional to  $\exp(v_{BE}/(2n_FV_t))$  instead of  $\exp(v_{BE}/(n_FV_t))$ . Then we find

$$HD_2 ext{ (high injection)} = \frac{V_{in}}{8n_F V_t}$$
 (8.21)

$$HD_3$$
 (high injection) =  $\frac{V_{in}^2}{96n_F^2V_t^2}$  (8.22)

It is seen that both under low- and high-injection conditions, the harmonic distortion is independent of technological parameters and independent of bias conditions.

Next we determine the intercept points for harmonic distortion under low-injection conditions. Using equations (2.15) and (2.16) we find for the intercept points  $IP_{2h}$  and  $IP_{3h}$ 

$$IP_{2h} = 4n_F V_t \tag{8.23}$$

$$IP_{3h} = 2\sqrt{6}\,n_F V_t \tag{8.24}$$

At room temperature and with  $n_F=1$ , one finds  $IP_{2h}\approx 103mV$  and  $IP_{3h}\approx 126mV$ .

When the input consists of two sine waves, then intermodulation products arise. Their value can be computed using equations (2.31) and (2.32) that are valid under low-distortion condition at low frequencies. In this way we find

$$IM_2 = \frac{V_{in}}{2n_E V_t} \tag{8.2}$$

$$IM_3 = \frac{V_{in}^2}{8n_F^2 V_t^2} \tag{8.24}$$

and consequently

$$IP_{2i} = 2n_F V_t \tag{8.27}$$

$$IP_{3i} = \sqrt{8} \, n_F V_t \tag{8.28}$$

At room temperature and with  $n_F = 1$  we find  $IP_{2i} \approx 50 mV$  and  $IP_{3i} \approx 73 mV$ .

#### 8.2.1.2 Computation of harmonics with the method of Section 5.3

The computation of the harmonics in the circuit of Figure 8.1 with the elementary transistor model can be performed without the knowledge of the calculation method of Section 5.3. It is instructive, however, to use this method to compute the harmonics of this simple circuit. For this method we start from the nonlinear circuit of Figure 8.2 that is equivalent to the circuit of Figure 8.1 for AC signals. This equivalent circuit contains two basic nonlinearities, a nonlinear

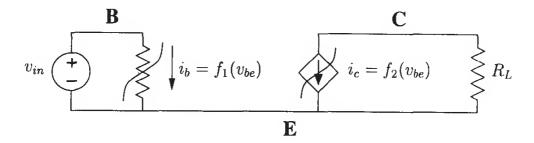


Figure 8.2: AC-equivalent circuit of the single-transistor amplifier of Figure 8.1 with a zero source resistance and including the nonlinear dependence of the base current and the collector current on the base-emitter voltage.

conductance, corresponding to the nonlinear base current and a nonlinear transconductance corresponding to the nonlinear collector current. The nonlinear conductance is placed in parallel with the input voltage source and can be discarded. The nonlinear transconductance is described by the relationship

$$i_c = g_m v_{be} + K_{2q_m} v_{be}^2 + K_{3q_m} v_{be}^3 + \dots$$
(8.29)

with (see also Section 6.2)

$$g_m = \frac{I_C}{V_t} \tag{8.30}$$

$$K_{2g_m} = \frac{I_C}{2!V_t^2} = \frac{g_m}{2V_t} \tag{8.31}$$

$$K_{3g_m} = \frac{I_C}{3!V_t^3} = \frac{g_m}{6V_t^2} \tag{8.32}$$

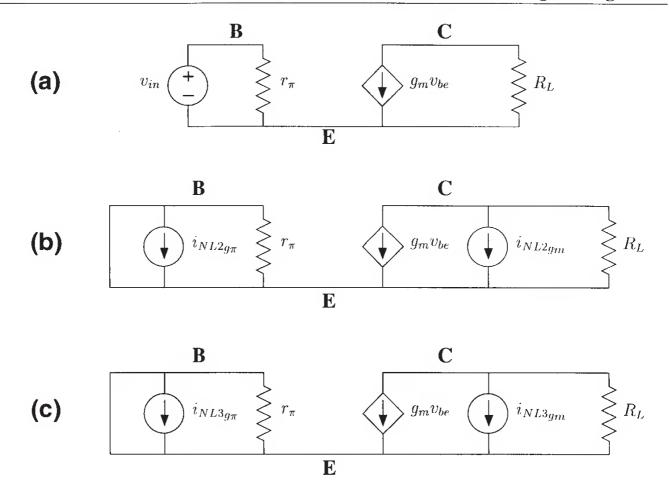


Figure 8.3: Linearized equivalent circuit of the single-transistor amplifier of Figure 8.2 excited with the appropriate inputs for the computation of the first-order response (a), the second-order response (b), and the third-order response (c).

Computation of the fundamental response First, the response of the linearized circuit is computed. This linearized circuit is shown in Figure 8.3a. From this circuit, one easily obtains the value of  $V_{out,1,0}$ 

$$V_{out,1,0} = -g_m R_L V_{in} (8.33)$$

which was also found in equation (8.14).

Computation of the second harmonic distortion The second harmonic at the output is found, by computing the output voltage of the linearized circuit in which the external excitation is removed while nonlinear current sources of order two are applied (see Figure 8.3b). The nonlinear second-order current source that corresponds to the nonlinearity of the base current is shortened and it does not play a role. The nonlinear current source of order two that corresponds to the collector current's nonlinearity is found from Table 5.5:

$$i_{NL2gm} = \frac{K_{2g_m}}{2} V_{be,1,0}^2 = \frac{K_{2g_m}}{2} V_{in}^2$$
(8.34)

The voltage over  $R_L$  in the circuit of Figure 8.3b yields the second harmonic of the output voltage. It is found to be

$$V_{out,2,0} = -R_L \cdot i_{NL2q_m} \tag{8.35}$$

Combining equations (8.33), (8.34) and (8.35) yields

$$HD_2 = \left| \frac{V_{out,2,0}}{V_{out,1,0}} \right| = \frac{K_{2g_m}}{2g_m} V_{in} = \frac{K'_{2g_m}}{2} V_{in}$$
 (8.36)

Using equation (8.31) we find again

$$HD_2 = \frac{V_{in}}{4V_t} \tag{8.37}$$

which is the same value as given in equation (8.17). Equation (8.36) illustrates the usefulness of working with a normalized nonlinearity coefficient: it is seen that the second harmonic distortion is simply proportional to the input amplitude and to the half of the second-order normalized coefficient. This is a general rule when only one nonlinearity is taken into account.

Computation of the third harmonic distortion The third harmonic is computed by computing the output voltage in the schematic of Figure 8.3c. The value of the nonlinear current source of order three is determined using Table 5.7:

$$i_{NL3gm} = K_{2g_m} V_{be,1,0} V_{be,2,0} + \frac{1}{4} K_{3g_m} V_{be,1,0}^3$$
(8.38)

Since  $v_{BE}$  is fixed at  $v_{IN}$ , which is a pure sine wave, all components of  $v_{BE}$  of order higher than one, are zero. Hence, the first term in equation (8.38) vanishes. Further we find

$$V_{out,3,0} = -R_L \cdot i_{NL3gm} \tag{8.39}$$

Combining equations (8.33), (8.38) and (8.39) we find

$$HD_3 = \left| \frac{V_{out,3,0}}{V_{out,1,0}} \right| = \frac{K_{3g_m}}{4g_m} V_{in}^2 = \frac{K'_{3g_m}}{4} V_{in}^2$$
 (8.40)

This equation reveals that the third harmonic distortion is proportional to the square of the input amplitude and to one fourth of the third-order normalized nonlinearity coefficient. Using equation (8.32) we find again

$$HD_3 = \frac{V_{in}^2}{24V_t^2} \tag{8.41}$$

which is the same value as given in equation (8.18).

## 8.2.2 Influence of the output resistance

The simple transistor model used in the previous sections is now extended with an Early resistance. If the Early effect is assumed to be linear, then the collector current is given by (see equation (6.2))

$$i_C = I_S \exp(\frac{v_{BE}}{V_t}) \left( 1 + \frac{v_{CE}}{V_{AF}} \right) \tag{8.42}$$

If the Early resistance is nonlinear, then the collector current is of the form

$$i_C = I_S \exp(\frac{v_{BE}}{V_t}) \cdot g(v_{CE}) \tag{8.43}$$

in which  $g(v_{CE})$  is a nonlinear function of  $v_{CE}$  only, as discussed in Section 6.2.2. With such model equation, it is complicated or even impossible to obtain an analytic expression for the input-output relation without imposing the restriction of weakly nonlinear behavior. However, with the calculation methods of either Section 5.2 or 5.3 it is still possible to obtain closed-form expressions for the different harmonics.

We will calculate the harmonics of the current that flows into the load resistance  $R_L$ . For the computation of harmonics with the calculation method of Section 5.3, we start from the equivalent circuit of Figure 8.4.

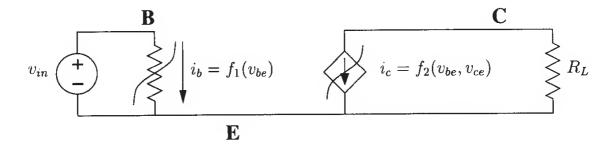


Figure 8.4: AC-equivalent circuit of the single-transistor amplifier of Figure 8.1 ( $R_S = 0\Omega$ ) with a nonlinear base current and a nonlinear collector current, which is now a function of two voltages.

The collector current is now a function of two voltages that can be expanded into a two-dimensional power series. The nonlinearity coefficients that are used in this power series have been discussed in Section 6.2.

#### 8.2.2.1 Fundamental response

First, the linearized equivalent circuit is analyzed. This is shown in Figure 8.5a. The output of interest is the current through the load resistance. It is easy to find that the first-order component of this current is given by

$$I_{out,1,0} = -\frac{g_m G_L}{G_L + g_o} V_{in} (8.44)$$

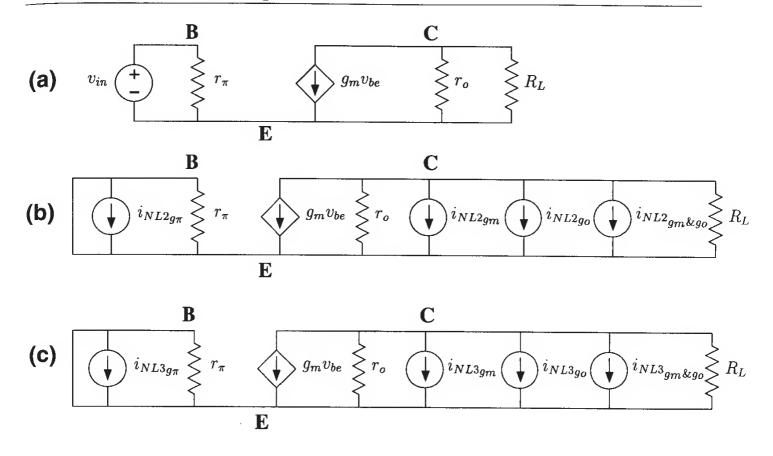


Figure 8.5: Linearized equivalent of the single-transistor amplifier of Figure 8.4 with appropriate excitations: (a) for the computation of the first-order response, (b) for the second-order response and (c) for the third-order response.

in which  $G_L = 1/R_L$  is the load conductance and  $g_o = 1/r_o$  is the output conductance.

#### 8.2.2.2 Second harmonic

Next, the second harmonic of the output current is computed. To this end, the linearized circuit is excited with four nonlinear current sources of order two, as shown in Figure 8.5b:  $i_{NL2g\pi}$  corresponds to the nonlinearity of the base current. Just as in the previous section, this current source does not play a role since it is shortened. The other three nonlinear current sources are  $i_{NL2gm}$ ,  $i_{NL2go}$  and  $i_{NL2gm\&go}$ . They correspond to the second-order nonlinearity coefficients  $K_{2gm}$ ,  $K_{2go}$  and  $K_{2gm\&go}$ , respectively. The value of the source  $i_{NL2gm}$  is the same as in the previous section. The sources  $i_{NL2go}$  and  $i_{NL2go}$  and  $i_{NL2go}$  are found using Table 5.5:

$$i_{NL2g_0} = \frac{K_{2g_0}}{2} V_{ce,1,0}^2 \tag{8.45}$$

$$=\frac{K_{2g_o}}{2} \left(\frac{g_m}{G_L + g_o}\right)^2 V_{in}^2 \tag{8.46}$$

and

$$i_{NL2_{gm}\&g_o} = \frac{K_{2_{g_m}\&g_o}}{2} V_{be,1,0} V_{ce,1,0} = -\frac{K_{2_{g_m}\&g_o}}{2} \frac{g_m}{G_L + g_o} V_{in}^2$$
(8.47)

The second harmonic of the output current is then given by

$$I_{out,2,0} = -\left(i_{NL2g_m} + i_{NL2g_m\&g_o} + i_{NL2g_o}\right) \frac{G_L}{G_L + g_o}$$
(8.48)

$$= -\left(K_{2g_m} - K_{2g_m \& g_o} \frac{g_m}{G_L + g_o} + K_{2g_o} \left(\frac{g_m}{G_L + g_o}\right)^2\right) \frac{G_L}{G_L + g_o} \frac{V_{in}^2}{2}$$
(8.49)

The factor  $g_m/(G_L+g_o)$  is the low-frequency voltage gain  $A_v$  of the single-transistor amplifier and hence

$$I_{out,2,0} = -\left(K_{2g_m} - K_{2g_m \& g_o} A_v + K_{2g_o} A_v^2\right) \frac{G_L}{G_L + g_o} \frac{V_{in}^2}{2}$$
(8.50)

and for the second harmonic of the output voltage we find

$$V_{ce,2,0} = -\left(K_{2g_m} - K_{2g_m \& g_o} A_v + K_{2g_o} A_v^2\right) \frac{1}{G_L + g_o} \frac{V_{in}^2}{2}$$
(8.51)

It is seen that  $K_{2g_m\&g_o}$  contributes a factor  $A_v$  more to the second harmonic than  $K_{2g_m}$ . The coefficient  $K_{2g_o}$  contributes a factor  $A_v^2$  more to the second harmonic than  $K_{2g_m}$ . On the other hand, the coefficients  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  are much smaller than  $K_{2g_m}$ . Equation (8.50) reveals that when  $R_L$  is a large resistance, such that  $A_v$  is high and the

Equation (8.50) reveals that when  $R_L$  is a large resistance, such that  $A_v$  is high and the collector of transistor  $Q_1$  is a high-impedance node, then the contributions of  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  might be considerable. This situation occurs for example when transistor  $Q_1$  is at the output of the first stage of a two-stage operational amplifier with Miller compensation. On the other hand, when  $R_L$  is low, then the contribution of  $K_{2g_m}$  is dominant. In the limit, when  $R_L$  is a short circuit, then the contributions of  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  vanish.

circuit, then the contributions of  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  vanish.

The result obtained in equation (8.50) can be intuitively explained as follows: when the voltage gain is high, then the voltage swing at the output is high. In this case, nonlinearities that are determined by the output voltage contribute much to the total harmonic.

In equation (8.50) the coefficients  $K_{2g_m}$ ,  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  can be replaced by their value given in Table 6.1. This yields:

$$I_{out,2,0} = -\left(\frac{g_m}{2V_t} - \frac{g_m}{V_{AF}}A_v + K_{2g_o}A_v^2\right)\frac{G_L}{G_L + g_o}\frac{V_{in}^2}{2}$$
(8.52)

Assume now that the Early effect is perfectly linear. Then the coefficient  $K_{2g_o}$  is zero, and the second harmonic of the output current is given by

$$I_{out,2,0} = -\left(\frac{g_m}{2V_t} - \frac{g_m}{V_{AF}}A_v\right) \frac{G_L}{G_L + g_o} \frac{V_{in}^2}{2}$$
(8.53)

Still, equation (8.53) contains a term that depends on the Early voltage. This means that the presence of the Early resistance, even when it is assumed to be perfectly linear, yields an additional distortion term, namely the second term of equation (8.53).

Numerical example We now evaluate the different terms of the second harmonic of the current through  $R_L$  as given in equation (8.50). To this purpose, we adopt the numerical values of the transistor parameters that have been used in Section 6.2.2 and we sweep the value of  $R_L$ . In this way, we have for a collector current of 0.85mA and  $g_m = 33mA/V$ . Further we have  $V_{AF} = 30V$  and  $V_{CE} = 2V$ . We now keep the collector-emitter voltage constant and we sweep the value of  $R_L$ . This means that the power supply voltage is adapted accordingly in order to keep  $v_{CE}$  constant. This is a rather artificial experiment but it provides useful insights.

The second harmonic of the current through  $R_L$  and its different contributions are shown in Figure 8.6 as a function of  $R_L$ . The value of the voltage gain  $A_v$  has been added to the x-axis.

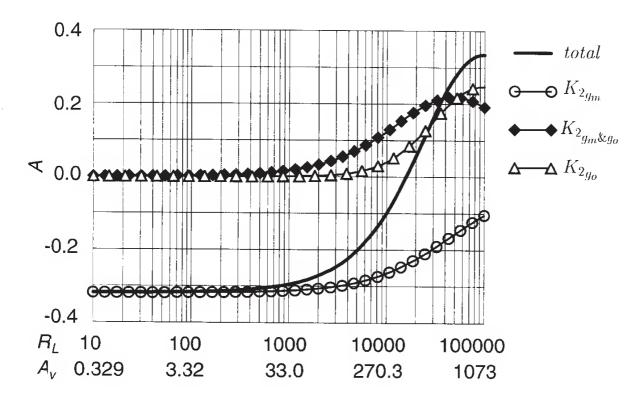


Figure 8.6: Second harmonic of the current through  $R_L$  and its different contributions according to equation (8.53). The input amplitude has been taken equal to 1V for reference.

The amplitude of the input voltage source has been taken equal to 1V for reference. In this way, the second harmonic of the current through  $R_L$  is artificially high.

It is seen that for low values of  $R_L$ , corresponding to a small voltage gain, the contribution of  $K_{2g_m}$  is dominant. This contribution remains dominant for a voltage gain of more than 100. At very high voltage gains the contributions of  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  become dominant. Such situation only occurs if a very high gain is realized in one stage, for example in operational transconductance amplifiers at low frequencies.

#### 8.2.2.3 Third harmonic

The third harmonic of the output current is computed using the circuit of Figure 8.5c. The value of the nonlinear current sources of order three can be found according to Table 5.7:

$$i_{NL3gm} = \frac{K_{3g_m}}{4} V_{be,1,0}^3 \tag{8.54}$$

$$i_{NL3g_o} = \frac{K_{3g_o}}{4} V_{ce,1,0}^3 + K_{2g_o} V_{ce,1,0} V_{ce,2,0}$$
(8.55)

and

$$i_{NL3_{gm}\&g_o} = \frac{K_{2_{g_m}\&g_o}}{2} V_{be,2,0} V_{ce,1,0} + \frac{K_{2_{g_m}\&g_o}}{2} V_{be,1,0} V_{ce,2,0} + \frac{K_{3_{2g_m}\&g_o}}{4} V_{be,1,0}^2 V_{ce,1,0} + \frac{K_{3_{g_m}\&g_o}}{4} V_{be,1,0}^2 V_{ce,1,0}$$

$$+ \frac{K_{3_{g_m}\&2g_o}}{4} V_{be,1,0} V_{ce,1,0}^2$$

$$(8.56)$$

Since  $V_{be,2,0} = 0$ , the first term of  $i_{NL3}{g_m\&g_o}$  is zero. The total third harmonic is the sum of the contributions of the above three nonlinear current sources. Since they are parallel to each other, the transfer function from one of these sources to the output is the same for the three sources. Using simple network analysis we find

$$I_{out,3,0} = -\left(i_{NL3g_m} + i_{NL3g_o} + i_{NL3g_m\&g_o}\right) \frac{G_L}{G_L + q_o} \tag{8.57}$$

Substituting the values of the nonlinear current sources into this expression and using the value of  $V_{ce,2,0}$  from equation (8.51) yields

$$I_{out,3,0} = -\frac{V_{in}^{3}}{4} \frac{G_{L}}{G_{L} + g_{o}} \left( K_{3g_{m}} + A_{v}^{2} K_{3g_{m} \& 2g_{o}} - A_{v}^{3} K_{3g_{o}} - A_{v} K_{32g_{m} \& g_{o}} \right.$$

$$+ \frac{A_{v}}{G_{L} + g_{o}} \left( K_{2g_{m} \& g_{o}} \right)^{2} - \frac{K_{2g_{m} \& g_{o}} K_{2g_{m}}}{G_{L} + g_{o}} + \frac{2A_{v}^{3}}{G_{L} + g_{o}} K_{2g_{o}}^{2}$$

$$+ \frac{2A_{v}}{G_{L} + g_{o}} K_{2g_{m}} K_{2g_{o}} - 3 \frac{A_{v}^{2}}{G_{L} + g_{o}} K_{2g_{m} \& g_{o}} K_{2g_{o}} \right)$$

$$(8.58)$$

The terms on the first line of this equation arise from third-order nonlinearities (third-order nonlinearity coefficients) whereas the next terms are caused by second-order nonlinearities. These produce a third-order signal by combining a second-order signal with a first-order one. Similar conclusions can be drawn from equation (8.58) as from the expression of the second harmonic when the collector is a low-impedance point, which means that  $R_L$  and  $A_v$  are low, then only  $K_{3g_m}$  gives a considerable contribution. In the case of a large value for  $R_L$  other contribution become important as well.

#### 8.2.3 Influence of the source resistance

When the voltage source that drives the single-transistor amplifier has an output resistance  $R_S$ , then the base-emitter voltage of transistor  $Q_1$  is no longer fixed to  $v_{IN}$ . As a result, the nonlinearity of the base current will play a role, which was not the case in the previous sections.

The circuit that is analyzed is given in Figure 8.7. The output of interest is the current through

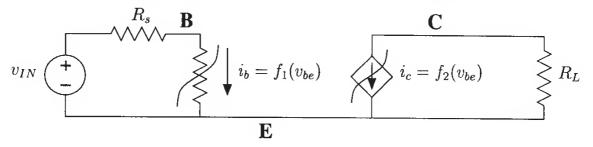


Figure 8.7: AC-equivalent circuit of the single-transistor amplifier of Figure 8.1 with a nonlinear base current and a nonlinear collector current and with  $R_S \neq 0\Omega$ .

the load resistance. For simplicity we neglect the Early effect.

#### 8.2.3.1 Fundamental response

The fundamental of the output current is computed from Figure 8.8a. The output current is given by the product of  $g_m$  with the fraction  $r_{\pi}/(R_S + r_{\pi})$  of  $V_{in}$ :

$$I_{out,1,0} = -g_m \cdot \frac{r_\pi}{R_S + r_\pi} \cdot V_{in}$$
 (8.59)

or in terms of conductances

$$I_{out,1,0} = -\frac{g_m G_S}{G_S + q_\pi} V_{in} \tag{8.60}$$

#### 8.2.3.2 Second harmonic

The second harmonic is computed using the circuit of Figure 8.8b. One easily finds

$$I_{out,2,0} = i_{NL2g_{\pi}} \frac{g_m}{G_S + q_{\pi}} - i_{NL2g_m}$$
 (8.61)

in which

$$i_{NL2g_{\pi}} = \frac{K_{2g_{\pi}}}{2} V_{be,1,0}^2 = \frac{K_{2g_{\pi}}}{2} \frac{G_S^2}{(G_S + g_{\pi})^2} V_{in}^2$$
(8.62)

$$i_{NL2gm} = \frac{K_{2g_m}}{2} V_{be,1,0}^2 = \frac{K_{2g_m}}{2} \frac{G_S^2}{(G_S + g_\pi)^2} V_{in}^2$$
(8.63)

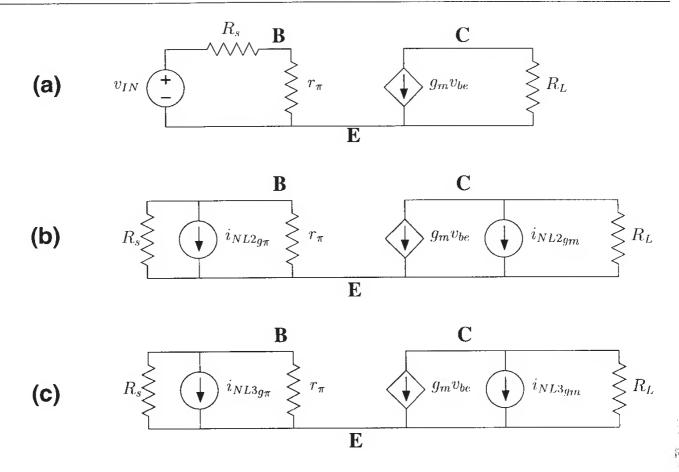


Figure 8.8: Linearized equivalent circuit of the circuit from Figure 8.7 used for the computation of the first-order response (a), for the computation of the second-order (b) and third-order response (c).

Substitution of equations (8.62) and (8.63) into equation (8.61) yields

$$I_{out,2,0} = \frac{G_S^2}{(G_S + g_\pi)^2} \left( K_{2g_\pi} \frac{g_m}{G_S + g_\pi} - K_{2g_m} \right) \frac{V_{in}^2}{2}$$
 (8.64)

It is seen in equation (8.64) that the contributions of the nonlinear collector current and of the nonlinear base current have an opposite sign. This means that they can cancel, either partially or completely, depending on the relative value of the two contributions. Also note that, when  $R_S$  goes to zero ( $G_S$  goes to infinity) then the contribution of the base current nonlinearity vanishes and we have again the situation of Section 8.2.1.

Under low injection conditions we find from Chapter 6 that the second-order nonlinearity coefficients  $K_{2g_m}$  and  $K_{2g_{\pi}}$  differ a factor equal to the AC beta, which is the ratio  $g_m/g_{\pi}$ :

$$\frac{K_{2g_m}}{K_{2g_\pi}} \approx \beta_{AC} = \frac{g_m}{g_\pi} \tag{8.65}$$

With this in mind we find from equation (8.64) that for low source resistances the contribution of the collector current nonlinearity (coefficient  $K_{2g_m}$ ) is dominant over the contribution of the

base current nonlinearity (coefficient  $K_{2g_{\pi}}$ ). In this case the second harmonic distortion is found from equations (8.60) and (8.64) to be

$$HD_2 \approx \frac{G_S}{G_S + g_\pi} \frac{K_{2g_m}}{2g_m} V_{in} = \frac{K'_{2g_m}}{2} \frac{r_\pi}{R_S + r_\pi} V_{in}$$
 (8.66)

The interpretation of this expression is as follows: a fraction  $[r_{\pi}/(R_S + r_{\pi})]$  of the input voltage is squared by the nonlinear transconductance, yielding a second-order component in the collector current.

Current source excitation Assume now that the source resistance  $R_S$  is much higher than  $r_{\pi}$ . This occurs when transistor  $Q_1$  is driven by a current source, rather than a voltage source, as shown in Figure 8.9. With  $G_S \ll g_{\pi}$ , and using  $V_{in} = I_{in}/G_S$  with  $I_{in}$  being the amplitude of

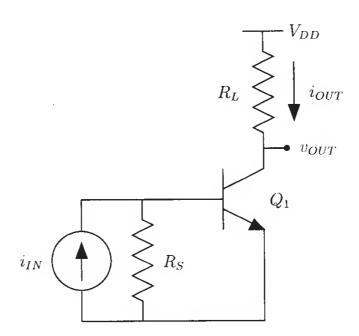


Figure 8.9: A one transistor amplifier (BJT version) with a current source excitation.

the input current, equation (8.64) reduces to

$$I_{out,2,0} = \frac{I_{in}^2}{2g_{\pi}^2} \left( K_{2g_{\pi}} \frac{g_m}{g_{\pi}} - K_{2g_m} \right)$$
 (8.67)

It is seen that the second harmonic of the output current is zero if

$$\frac{g_m}{g_\pi} = \frac{K_{2g_m}}{K_{2g_\pi}} \tag{8.68}$$

This condition is satisfied if the transistor beta is bias independent such that the base current nonlinearity and the collector current nonlinearity track. Indeed, in that case the collector current

and the base current differ by a constant  $\beta_F$ , as well as their derivatives. Then the ratios on both sides of equation (8.68) are equal to  $\beta_F$ . It is clear that in this case no distortion occurs: the input current is amplified by a constant factor  $\beta_F$ . This perfectly linear behavior can also be interpreted as a case of pre- and post-distortion as discussed in Section 4.7: the input current is transformed by the base-emitter diode into a voltage. This transformation is nonlinear. Next, the resulting voltage determines the collector current. The relation between the collector current and the base-emitter voltage is again nonlinear and it is seen that the relationship between the input current and the base-emitter voltage is the same, apart from a constant factor, as the relationship between the collector current and the controlling base-emitter voltage. Since we are using the latter relationship in the reverse direction, no distortion results.

#### 8.2.3.3 Third harmonic

The third harmonic of the output current is found in a way that is similar to the computation of the second harmonic:

$$I_{out,3,0} = i_{NL3g\pi} \frac{g_m}{G_S + g_\pi} - i_{NL3gm}$$
 (8.69)

as can be seen from Figure 8.8c. The nonlinear current sources of order three in this equation are given by

$$i_{NL3g_{\pi}} = \frac{K_{3g_{\pi}}}{4} V_{be,1,0}^3 + K_{2g_{\pi}} V_{be,1,0} V_{be,2,0}$$
(8.70)

$$i_{NL3gm} = \frac{K_{3g_m}}{4} V_{be,1,0}^3 + K_{2g_m} V_{be,1,0} V_{be,2,0}$$
(8.71)

It is seen that a knowledge of the second harmonic of the base-emitter voltage is required for the computation of the third harmonic of the output current. This second harmonic can be computed from the circuit of Figure 8.8b. It is given by

$$V_{be,2,0} = -i_{NL2g_{\pi}} \cdot (G_S + g_{\pi}) \tag{8.72}$$

Using equation (8.62) this becomes

$$V_{be,2,0} = -\frac{K_{2g_{\pi}}}{2} \frac{G_S^3}{(G_S + g_{\pi})^3} \cdot V_{in}^2$$
(8.73)

Combining equations (8.73), (8.69), (8.70), (8.71) and (8.72) and yields

$$I_{out,3,0} = \frac{V_{in}^3 G_S^3}{4 (G_S + g_\pi)^5} \left( -(g_\pi + G_S)^2 K_{3g_m} + g_m (g_\pi + G_S) K_{3g_\pi} + 2 (g_\pi + G_S) K_{2g_m} K_{2g_\pi} - 2g_m K_{2g_\pi}^2 \right)$$

$$(8.74)$$

Under low-injection conditions the coefficients  $K_{2g_m}$  and  $K_{2g_\pi}$  approximately differ by a factor  $\beta_{AC}$  and the same holds for the coefficients  $K_{3g_m}$  and  $K_{3g_\pi}$ . Then when the source resistance

is low ( $G_S$  is high) the first term inside the brackets of equation (8.74) is dominant over the three other terms. In this case the third harmonic distortion is found from equations (8.60) and (8.74):

$$HD_3 = \frac{G_S^2}{(G_S + g_\pi)^2} \frac{K_{3g_m}}{4g_m} V_{in}^2 = \frac{K_{3g_m}'}{4} \frac{r_\pi^2}{(R_S + r_\pi)^2} V_{in}^2$$
(8.75)

The interpretation is similar to  $HD_2$ : a fraction  $[r_{\pi}/(R_S + r_{\pi})]$  of the input voltage is raised to the third power by the third-order nonlinearity of the transconductance, yielding a third-order component in the collector current.

Current source excitation Assume again that the transistor is driven by a current, such that  $G_S \ll g_{\pi}$ . Using  $G_S V_{in} = I_{in}$  the third harmonic  $I_{out,3,0}$  given in equation (8.74) reduces to

$$I_{out,3,0} = \frac{I_{in}^3}{4g_{\pi}^5} \left( -g_{\pi}^2 K_{3g_m} + g_m g_{\pi} K_{3g_{\pi}} + 2g_{\pi} K_{2g_m} K_{2g_{\pi}} - 2g_m K_{2g_{\pi}}^2 \right)$$
(8.76)

Assume again that the transistor beta is constant such that the base current nonlinearity and the collector current nonlinearity track. In this case

$$\frac{g_m}{g_\pi} = \frac{K_{2g_m}}{K_{2g_\pi}} = \frac{K_{3g_m}}{K_{3g_\pi}} = \beta_F \tag{8.77}$$

and it is easy to verify with equation (8.76) that  $I_{out,3,0}$  then again reduces to zero. From this, one can conclude that a bipolar transistor with a constant beta does not yield distortion when it is driven by a current source.

#### 8.2.4 Influence of the base resistance

As explained in Section 6.4, the base resistance  $r_B$  falls apart in two parts: the extrinsic part  $r_{Bex}$ , which is assumed to be linear, and the intrinsic part  $r_{Bi}$  which depends on the base current. Instead of considering the nonlinearity of the base resistance as a current-controlled nonlinearity, we will make the calculations in terms of a voltage-controlled nonlinearity. In this case, the current through the base resistance is a function of the voltage over the intrinsic base resistance. This approach will ease the calculations. It is allowed since the nonlinear relationship between the voltage over the intrinsic base resistance and the base current can be inverted. Further, the extrinsic base resistance is joined with the series resistance  $R_S$  of the voltage source.

The nonlinear circuit that is analyzed is shown in Figure 8.10. It contains three nonlinearities: the nonlinear base current, which is described with the relationship  $i_b = f_1(v_{be})$ , the nonlinear collector current, described as  $i_c = f_2(v_{be})$  and the nonlinear intrinsic base resistance, described with the function  $i_b = f_3(v_{b''b})$ . The node **B**" is the node between the intrinsic and the extrinsic base resistance. The output of interest is the AC short-circuit current between the collector and emitter. By shorting the output in AC, the output conductance does not play any role in the analysis here.

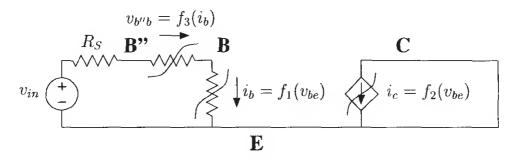


Figure 8.10: AC-equivalent nonlinear circuit of the single-transistor amplifier including a non-linear intrinsic base resistance. The extrinsic base resistance is added to the source resistance. The output of interest is the AC short-circuit current at the output.

#### 8.2.4.1 Fundamental response

The fundamental of the output current is computed in the same way as in Section 8.2.3. This yields

$$I_{out,1,0} = -\frac{g_m G_S'}{g_\pi + G_S'} V_{in} = \frac{g_m r_\pi}{r_\pi + R_S'} V_{in}$$
(8.78)

in which  $G'_S = 1/R'_S$  and  $R'_S = R_S + r_{Bi}$ .

#### 8.2.4.2 Second harmonic

For the computation of the second harmonic, three nonlinear current sources are taken into account, each corresponding to one nonlinearity. This yields the circuit shown in Figure 8.11. The

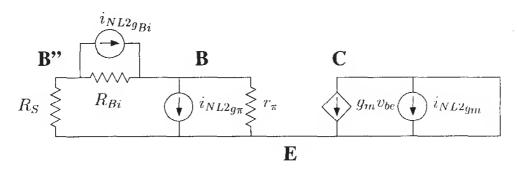


Figure 8.11: Circuit to be solved for the computation of the second harmonic of the AC short circuit output current, including the effect of a nonlinear base resistance.

contributions to the overall second harmonic of the three nonlinear current sources are given by

$$I_{out,2,0a} = i_{NL2g_m} \cdot TF_{i_{NL2g_m} \to output}$$

$$= K_{2g_m} \left(\frac{r_{\pi}}{r_{\pi} + R_S'}\right)^2 \frac{V_{in}^2}{2} \cdot (-1)$$
(8.79)

which is the contribution of the nonlinearity of the transistor collector current,

$$I_{out,2,0b} = i_{NL2g_{\pi}} \cdot TF_{i_{NL2g_{\pi}} \to output}$$

$$= K_{2g_{\pi}} \left(\frac{r_{\pi}}{r_{\pi} + R'_{S}}\right)^{2} \frac{V_{in}^{2}}{2} g_{m} \left(r_{\pi} \parallel R'_{S}\right)$$
(8.80)

which is the contribution of the nonlinearity of the transistor base current. Finally, the contribution of the nonlinearity of the base resistance is given by

$$I_{out,2,0c} = i_{NL2g_{Bi}} \cdot TF_{i_{NL2g_{Bi}} \to output}$$

$$= K_{2g_{Bi}} \left(\frac{r_{Bi}}{r_{\pi} + R'_{S}}\right)^{2} \frac{V_{in}^{2}}{2} \beta_{AC} \frac{r_{Bi}}{r_{\pi} + R'_{S}}$$
(8.81)

The total second harmonic is the sum of  $I_{out,2,a}$ ,  $I_{out,2,b}$  and  $I_{out,2,c}$ . It is seen that the contribution of  $K_{2g_m}$  has a different sign from the contributions of  $K_{2g_{B_i}}$  and  $K_{2g_{\pi}}$ . This means that the different contributions can cancel, at least partially, as we saw in Section 8.2.3.2.

It is interesting to check whether the contribution of the nonlinearity of the base resistance is significant compared to the other two contributions. Equation (8.81) reveals that the base resistance contribution increases when the source resistance becomes smaller. When the source resistance is zero, then  $R_S$  is nothing else but  $r_{Bex}$ . In this case, the ratio of the base resistance contribution to the collector current contribution can be computed from equations (8.79) and (8.81). This yields

$$\frac{\text{contribution of collector current}}{\text{contribution of base resistance}} = -\frac{K_{2g_m}}{K_{2g_{Bi}}} \cdot \frac{r_\pi^2}{r_{Bi}^2} \cdot \frac{r_\pi + r_{Bi} + r_{Bex}}{\beta_{AC} \cdot r_{Bi}}$$
(8.82)

It is seen that the contribution of the base resistance nonlinearity is small when the base resistance is small compared to  $r_{\pi}$  and when  $K_{2g_{Bi}}$  is small.

The relative values of the three contributions are further examined by a numerical evaluation of the different contributions to the second harmonic. To this purpose, the contributions are evaluated as a function of the bias current using the SPICE parameters of Table 8.1.

The different contributions have been computed with the Gummel-Poon model equations for the collector and base current, equations (6.7) and (6.18), respectively, and with the power model for the intrinsic base resistance, equation (6.30). The different contributions and the total second harmonic are shown in Figure 8.12 as a function of the collector current. The contributions are normalized to the fundamental response (equation (8.78)), yielding the second harmonic distortion. The collector current has been normalized to  $I_{KF}$ , the forward knee current.

It is seen that both under low-injection (collector current far below  $I_{KF}$ ) and high-injection conditions (collector current far above  $I_{KF}$ ), the nonlinearity of the base resistance can be neglected for the computation of the second harmonic at low frequencies. In the vicinity of  $I_{KF}$  the contribution of the base resistance nonlinearity is significant, because the contribution of the collector current nonlinearity cancels nearly completely with the contribution of the base current

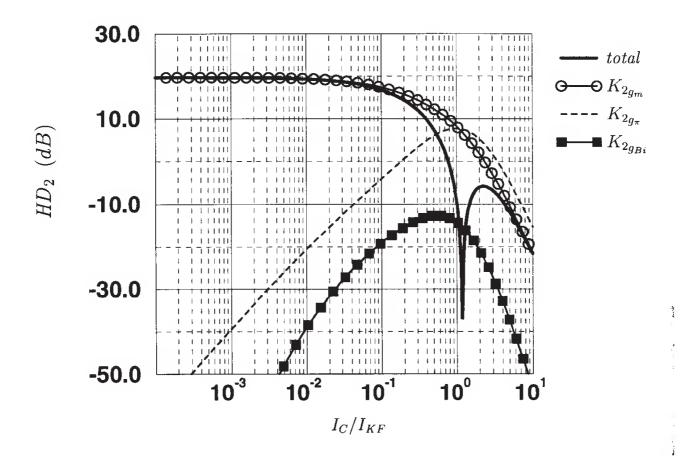


Figure 8.12: Second harmonic distortion and its different contributions as a function of the collector current of the transistor in the single-transistor amplifier of Figure 8.1 with a zero source resistance and a zero load resistance (AC short circuit). The model parameters used for these computations are given in Table 8.1.  $HD_2$  is normalized to an input amplitude of 1 Volt the collector current is normalized to  $I_{KF}$ .

nonlinearity. Further it is seen that at low-injection conditions, the second harmonic is nearly completely determined by the nonlinearity of the collector current.

The model for the nonlinearity of the base resistance is the power model that only model emitter crowding. Using the model of Chiu (see Section 6.4) a smaller value of the base resistance will result. This will yield an even smaller contribution of the base resistance nonlinearity at low to moderate currents. The collector current region in the vicinity of  $I_{KF}$  and higher is seldon used. Indeed, when a high collector current is required, then usually a transistor with a large emitter area is taken which has a higher  $I_{KF}$  and a smaller base resistance. In addition, the different effects that modify the base resistance and therefore lead to a nonlinear base resistance are postponed to larger currents as well.

Assume now that the transistor is driven by a current source. With  $R'_S \gg r_\pi$ , we find

$$\frac{\text{contribution of collector current}}{\text{contribution of base resistance}} = -\frac{K_{2g_m}}{K_{2g_{Bi}}} \cdot \frac{1}{\beta_{AC} \, r_{Bi} \, G_S'} \tag{8.83}$$

This means that for a current drive the base resistance does not play a role. This can be understood as follows: the input current flows through the base-emitter diode and gives rise to a base-emitter voltage. The current through the base-emitter diode does not change due to the presence of the base resistance, whether it is linear or not.

#### 8.2.4.3 Third harmonic

The computation of the third harmonic of the AC short-circuit current of Figure 8.10 yields an expression that is quite involved. The reason is that three nonlinearities are involved, each giving a contribution proportional to their second-order nonlinearity coefficient and a contribution proportional to their third-order nonlinearity coefficient (see Table 5.7). Hence we have six contributions to the third harmonic.

The six contributions to the overall third harmonic distortion have been computed as a function of the collector current using the accurate model equations for the collector current (equation (6.7)) and the base current (equation 6.18) and the power model for the base resistance (equation (6.30)). Again, the SPICE model parameters of Table 8.1 have been used. The different contributions are shown in Figure 8.13. The collector current has again been normalized with to  $I_{KF}$ , and the third harmonic distortion to an input amplitude of 1V. First, the contributions of third-order nonlinearity coefficients are studied. It is seen that for currents well below  $I_{KF}$ the third harmonic distortion is governed by the third-order nonlinearity of the collector current (nonlinearity coefficient  $K_{3q_m}$ ). Second-order coefficients do not play a significant role at these low currents. This can be understood as follows: at low bias currents,  $r_{\pi} \gg r_{B}$ , such that the voltage drop over  $r_B$  is negligible and the voltage over  $r_{\pi}$  is almost equal to the input voltage. Since the input voltage source is assumed to be purely sinusoidal, the voltage over  $r_{\pi}$  does not contain higher harmonics. Hence, no voltage that determines a nonlinearity has a second-order component. In this way, the second-order nonlinearities do not produce a third-order signal. Hence, it can be concluded that at bias currents that are sufficiently smaller than  $I_{KF}$  the third harmonic distortion is given by expression (8.18), which holds for a simple transistor model.

Consider now the contribution of the other two third-order nonlinearities  $K_{3g_{\pi}}$  and  $K_{3g_{Bi}}$ . The ratio of the contribution of  $K_{3g_{\pi}}$  to the contribution of  $K_{3g_{m}}$  can be derived from equation (8.74), in which  $G_{S}$  is replaced by  $G_{S}'$ :

$$\frac{\text{contribution of } K_{3g_{\pi}}}{\text{contribution of } K_{3g_m}} = -\frac{g_m K_{3g_{\pi}}}{(g_{\pi} + G'_S) K_{3g_m}}$$
(8.84)

At low currents and with a low source resistance this ratio reduces to

$$\frac{\text{contribution of } K_{3g_{\pi}}}{\text{contribution of } K_{3g_{m}}}\bigg|_{\text{low injection}} \approx -\frac{g_{m}K_{3g_{\pi}}}{G'_{S}K_{3g_{m}}} \approx -\frac{g_{m}R'_{S}}{\beta_{AC}} = -\frac{R'_{S}}{r_{\pi}} \tag{8.85}$$

The ratio in the right-hand side reduces to  $r_B/r_\pi$  when the source resistance is zero. This ratio is usually much smaller than one at low currents. For example, at a collector current of 1mA and a  $\beta_F$  of 120 (see Table 8.1),  $r_\pi$  is about  $3.1k\Omega$  while the base resistance at low currents is found from Table 8.1 to be  $378\Omega$ . The ratio of equation (8.85) is then 0.12. This low ratio is in agreement with the curves in Figure 8.13.

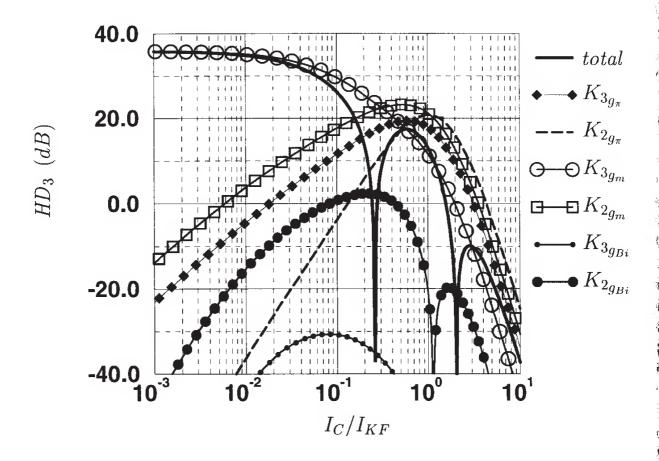


Figure 8.13: Third harmonic distortion and its different contributions as a function of the collector current of the transistor in the single-transistor amplifier of Figure 8.1 with a zero source resistance and a zero load resistance (AC short circuit). The model parameters used for these computations are given in Table 8.1.  $HD_3$  is normalized to an input amplitude of 1V, the collector current is normalized to  $I_{KF}$ .

At high collector currents this ratio can be computed using the value of  $K_{3g_m}$  from equation (6.10) and the value of  $K_{3g_m}$  from equation (6.27)

$$K_{3g_m}|_{\text{high injection}} = \frac{I_C}{48V_t^3}$$
 and  $K_{3g_\pi} = \frac{I_B}{6V_t^3}$  (8.86)

Further we have

$$g_m|_{\text{high injection}} = \frac{I_C}{2V_t}$$
 and  $g_\pi = \frac{I_B}{V_t}$  (8.87)

Using these values we find from equation (8.84)

$$\left. \frac{\text{contribution of } K_{3g_{\pi}}}{\text{contribution of } K_{3g_m}} \right|_{\text{high injection}} \approx -\frac{g_m K_{3g_{\pi}}}{g_{\pi} K_{3g_m}} \approx -16 \tag{8.88}$$

Hence, at very high currents the contribution of  $K_{3g_{\pi}}$  is sixteen times higher than the contribution of  $K_{3g_m}$ .

As seen in Figure 8.13, the contribution of  $K_{3g_{Bi}}$  is small in all bias regions. Hence it can be neglected, except at bias values where the largest contributions mutually cancel: then the total third harmonic distortion is small and small contributions become significant.

Next, consider the contributions of second-order nonlinearity coefficients. First, it is seen in Figure 8.13 that the contribution of  $K_{2g_{Bi}}$  is negligible except at the collector current region where the large contributions cancel. Next, it is seen that the contribution of the other second-order nonlinearities is insignificant at low currents. From the high-injection region on, the nonlinearity coefficients  $K_{2g_m}$  and  $K_{2g_\pi}$  play a significant role. This can be explained by the fact that a high currents,  $r_\pi$  is no longer much larger than  $r_B$ . In this case, we have a nonlinear voltage divider composed of  $r_B$  and  $r_\pi$  and the voltage over  $r_\pi$  not only has a first-order component but also a second-order one, as pointed out in Section 5.5.1. The second-order nonlinearities  $K_{2g_\pi}$  and  $K_{2g_m}$  combine the first- and the second-order components and each produce a third-order signal. Just as with the contributions of the third-order nonlinearity coefficients of  $r_\pi$  and  $r_B$  it is seen that at low bias currents the contribution of the second-order nonlinearity coefficient of  $g_m$  is much larger than the contribution of the second-order nonlinearity of  $r_\pi$ , whereas under high-injection conditions the contribution of  $K_{2g_\pi}$  is higher than the contribution of  $K_{2g_m}$ .

As in Section 8.2.4.2 it should be noticed that the collector current region in the vicinity of  $I_{KF}$  and higher is seldom used. If we only consider the current region at low values of  $i_C/I_{KF}$  then only the third-order nonlinearity of the collector current needs to be taken into account.

Finally, we can remark again that, just as in the case of the second harmonic, the base resistance does not influence the third harmonic when the transistor is driven by a current source.

## 8.2.5 Influence of $C_{\pi}$

The base-emitter capacitance  $C_{\pi}$  influences the nonlinear behavior of the single-transistor circuit in two ways: first, it is an additional nonlinear element. Next, even if  $C_{\pi}$  would be considered as purely linear, its presence will affect the values of the nonlinear current sources of the static nonlinearities, as well as the transfer functions from these current sources to the circuit output.

We will analyze the nonlinear behavior of the single-transistor with a zero AC load resistance. The output of interest is then the output current. Taking into account the presence of  $C_{\pi}$ , the distortion on the output current is analyzed in two cases: in the first case, the amplifier is driven by an AC current. In this case we can neglect the nonlinearity of the base resistance, as we saw in the previous section. Next, the amplifier is excited by a voltage source and the nonlinearity of the base resistance is taken into account.

#### 8.2.5.1 Current drive

The circuit that is analyzed is shown in Figure 8.14. The output of interest is the AC short-circuit current at the output. Three nonlinearities are taken into account:

- the nonlinear base current, which is described as  $i_b = f_1(v_{be})$ .

- the nonlinear collector current, described as  $i_c = f_2(v_{be})$ . The dependence on  $v_{ce}$  is not considered here, since the output conductance is shorted in AC.
- the nonlinearity of  $C_{\pi}$ . This nonlinearity is described by a relationship between the charge upon  $C_{\pi}$  and  $v_{be}$ .

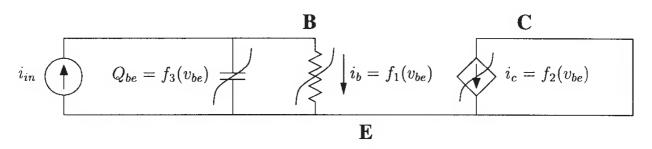


Figure 8.14: AC-equivalent circuit of the single-transistor amplifier of Figure 8.1 driven by current source and AC shorted at its output. The nonlinear capacitance  $C_{\pi}$  is included.

The fundamental of the AC short-circuit output current is given by

$$I_{out,1,0} = \frac{g_m}{g_\pi + sC_\pi} I_{in} = \frac{\beta_{AC} I_{in}}{1 + sC_\pi r_\pi}$$
(8.89)

The second harmonic is computed using the circuit of Figure 8.15. The value of the different

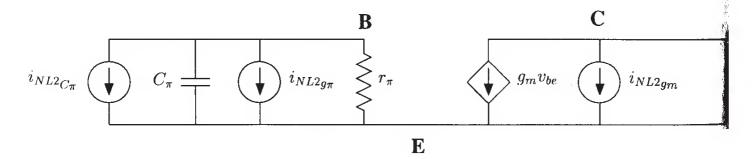


Figure 8.15: Circuit to be analyzed for the computation of the second harmonic of the AC short circuit output current of the single-transistor amplifier including  $C_{\pi}$ .

nonlinear current sources can be computed using Table 5.5. They are given by:

$$i_{NL2g_{\pi}} = \frac{K_{2g_{\pi}}}{2} V_{be,1,0}^2 = \frac{K_{2g_{\pi}}}{2} \frac{1}{(g_{\pi} + sC_{\pi})^2} I_{in}^2$$
(8.90)

$$i_{NL2_{C_{\pi}}} = sK_{2_{C_{\pi}}}V_{be,1,0}^2 = sK_{2_{C_{\pi}}}\frac{1}{(g_{\pi} + sC_{\pi})^2}I_{in}^2$$
 (8.9)

$$i_{NL2g_m} = \frac{K_{2g_m}}{2} V_{be,1,0}^2 = \frac{K_{2g_m}}{2} \frac{1}{(g_\pi + sC_\pi)^2} I_{in}^2$$
(8.92)

The second harmonic  $I_{out,2,0}$  of the output current is found by multiplying each nonlinear current source with its transfer function to the output and summing the contributions:

$$I_{out,2,0} = i_{NL2g_{\pi}} \cdot TF_{i_{NL2}g_{\pi} \to i_{out}} + i_{NL2g_{m}} \cdot TF_{i_{NL2}g_{m} \to i_{out}} + i_{NL2}C_{\pi} \cdot TF_{i_{NL2}C_{\pi} \to i_{out}}$$
(8.93)

Hereby we notice that, according to the theory developed in Section 5.3 the transfer functions must be evaluated for the frequency variable 2s instead of s. The transfer functions are found using simple network analysis. They are given by

$$TF_{i_{NL2}g_{\pi} \to i_{out}} = \frac{g_m}{g_{\pi} + 2s C_{\pi}} \tag{8.94}$$

$$TF_{i_{NL2g_m} \to i_{out}} = -1 \tag{8.95}$$

$$TF_{i_{NL_{2}C_{\pi}} \to i_{out}} = \frac{g_{m}}{g_{\pi} + 2s C_{\pi}}$$
 (8.96)

Hence we find for the second harmonic  $I_{out,2,0}$  of the output current

$$I_{out,2,0} = \left(i_{NL2g_{\pi}} + i_{NL2C_{\pi}}\right) \frac{g_m}{q_{\pi} + 2sC_{\pi}} - i_{NL2g_m}$$
(8.97)

Substituting equations (8.90) through (8.92) into equation (8.97) yields for the second harmonic of the current

$$I_{out,2,0} = \frac{g_m \left(2sK_{2C_{\pi}} + K_{2g_{\pi}}\right) - (g_{\pi} + 2sC_{\pi})K_{2g_m}}{2\left(g_{\pi} + sC_{\pi}\right)^2 \left(g_{\pi} + 2sC_{\pi}\right)}I_{in}^2$$
(8.98)

Assume now that the nonlinearity coefficients in the above equation are given by the first-order expressions (see Table 6.1 and (6.47)):

$$K_{2C_{\pi}} = \frac{\tau_F g_m}{2V_t} \qquad K_{2g_{\pi}} = \frac{g_{\pi}}{2V_t} \qquad K_{2g_m} = \frac{g_m}{2V_t}$$
 (8.99)

With these values the second harmonic of the output current in equation (8.98) vanishes. This is due to the fact that the three nonlinearities "track", as we defined in Section 3.2.6. Here we have a unique situation where a capacitive nonlinearity tracks with a conductive nonlinearity.

This tracking of nonlinearities yields a transistor beta which changes with frequency while it is independent of the bias current. In reality, however, the second harmonic is not zero because the nonlinearity coefficients do not satisfy the simple expressions listed above.

For the third harmonic, a similar conclusion can be drawn: if the first-order expressions for the different involved nonlinearity coefficients of order two and three are used, then one will find that the third harmonic is zero.

## 8.2.5.2 Voltage drive

The single-transistor amplifier is now excited with a voltage source. We will again analyze the AC short-circuit current at the output. Compared to the case where the circuit was driven

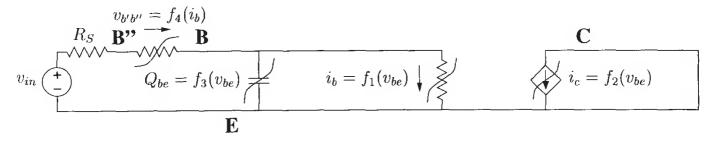


Figure 8.16: AC-equivalent nonlinear circuit of the single-transistor amplifier of Figure 8.1 driven by a voltage source and AC shorted at its output and including four nonlinearities.

by a current source, as in the previous section, we will now explicitly take into account the nonlinearity of the intrinsic base resistance. This is described with a function of the form  $i_b = f_4(v_{b''b})$ . Hence the circuit to be analyzed has four nonlinearities, as shown in Figure 8.16. The fundamental of the output current is found to be

$$I_{out,1,0} = -\frac{g_m G_S'}{g_\pi + G_S' + sC_\pi} V_{in}$$
 (8.100)

in which  $R'_S = R_S + r_{Bi}$  and  $G'_S = 1/R'_S$ . Note again that the extrinsic base resistance has been joined with the source resistance, resulting into the resistance  $R_S$ .

Analysis of the second harmonic distortion For the computation of the second harmonic, the linearized equivalent of the circuit of Figure 8.16 is excited by four nonlinear current sources of order two. We leave the detailed computations as an exercise to the reader. We will only present the results here.

The second harmonic consists of contributions from the nonlinear  $C_{\pi}$ , from the base resistance, from the nonlinear  $r_{\pi}$  (or  $g_{\pi}$ ) and from the nonlinear transconductance. It is given by

$$I_{out,2,0} = \frac{2sK_{2C_{\pi}} + K_{2g_{\pi}}}{2} \left( \frac{r_{\pi}}{(r_{\pi} + R'_{S}) (1 + sC_{\pi} (r_{\pi} \parallel R'_{S}))} V_{in} \right)^{2} \cdot \frac{g_{m} (r_{\pi} \parallel R'_{S})}{1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S})} + \frac{K_{2g_{m}}}{2} \left( \frac{r_{\pi}}{(r_{\pi} + R'_{S}) (1 + sC_{\pi} (r_{\pi} \parallel R'_{S}))} V_{in} \right)^{2} \cdot (-1) + \frac{K_{2g_{Bi}}}{2} \left( \frac{r_{Bi} (1 + sC_{\pi} r_{\pi})}{(r_{\pi} + R'_{S}) (1 + sC_{\pi} (r_{\pi} \parallel R'_{S}))} V_{in} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))} \right)^{2} \cdot \frac{\beta_{AC} r_{Bi}}{(r_{\pi} + R'_{S}) (1 + 2sC_{\pi} (r_{\pi} \parallel R'_{S}))}$$

Since  $r_{\pi}$  and  $C_{\pi}$  are parallel elements, their respective contributions only differ in the nonlinearity coefficients. The ratio of the contribution of the collector current nonlinearity to the sum of the contribution of the base current and  $C_{\pi}$  is given by

$$\frac{\text{contribution of collector current}}{\text{contribution of base current and } C_{\pi}} = -\frac{K_{2g_{m}}}{K_{2g_{\pi}} + 2sK_{2C_{\pi}}} \cdot \frac{1 + 2sC_{\pi}\left(r_{\pi} \parallel R_{S}'\right)}{g_{m}\left(r_{\pi} \parallel R_{S}'\right)} \quad (8.102)$$

Under low-injection conditions the different second-order nonlinearity coefficients in this ratio can be approximated well by their first-order expression:

$$K_{2g_m} = \frac{g_m}{2V_t}$$
  $K_{2g_\pi} = \frac{g_m}{2\beta_{AC}V_t}$   $K_{2C_\pi} = \frac{\tau_F g_m}{2V_t}$  (8.103)

This yields

$$\frac{\text{contribution of collector current}}{\text{contribution of base current and } C_{\pi}} = -\frac{\beta_{AC}}{1 + 2s\tau_{F}\beta_{AC}} \cdot \frac{1 + 2sC_{\pi}\left(r_{\pi} \parallel R'_{S}\right)}{g_{m}\left(r_{\pi} \parallel R'_{S}\right)} \tag{8.104}$$

This equation can be interpreted as follows. At low frequencies and with  $R_S' \ll r_{\pi}$ , the ratio of the two contributions is approximately  $r_{\pi}/R_S'$ . The ratio falls off with 20dB per decade from the frequency  $f_1$  that is given by

$$f_1 = \frac{1}{4\pi\beta_{AC}\tau_F} \approx \frac{1}{4\pi r_\pi C_\pi}$$
 (8.105)

This frequency is a factor  $2\beta_{AC}$  smaller than the transistor cutoff frequency  $f_T$ .

The ratio of the contributions is again constant from the frequency  $f_2$  on, which is given by

$$f_2 = \frac{1}{4\pi \left(r_\pi \parallel R_S'\right) C_\pi} \tag{8.106}$$

This frequency is usually much higher than  $f_1$  since  $R_S'$  is the sum of the low output resistance of the voltage source and the base resistance. If the source resistance is zero, then  $R_S' = r_{Bi}$ . For a collector current of 1mA and with the model parameters of Table 8.1 we find  $f_1 = 66MHz$  and  $f_2 = 609.8MHz$ .

Further, it is interesting to compare the contribution of the nonlinear base resistance to the other contributions, in the same way as in Section 8.2.4.2. Again, this comparison is made for a zero source resistance, since in this situation the contribution of the nonlinear base resistance has a larger influence. The ratio of the base resistance contribution and the collector current contribution as computed from equation (8.101) is given by

$$\frac{\text{contribution of collector current}}{\text{contribution of base resistance}} = -\frac{K_{2g_m}}{K_{2g_{Bi}}} \cdot \frac{r_\pi^2}{r_{Bi}^2} \cdot \frac{r_\pi + r_{Bi} + r_{Bex}}{\beta_{AC} r_{Bi}} \cdot \frac{(1 + 2sC_\pi \left(r_\pi \parallel r_B\right))}{(1 + sC_\pi r_\pi)^2}$$

$$(8.107)$$

The low-frequency value of this ratio has been computed in equation (8.82) and shown in Figure 8.12. From there we know that the contribution of the base resistance reaches a maximum at a collector current value close to  $I_{KF}$ . From equation (8.107) we now see that the ratio of the two contributions decreases with frequency at a rate of  $40 \, dB$  per decade from the frequency  $f_{\beta} = 1/(2\pi r_{\pi}C_{\pi})$ , due to a double pole at this frequency. At the frequency  $f_2$  (see equation (8.106)) a zero occurs, such that from  $f_2$  the ratio only decreases at a rate of  $20 \, dB$  per decade.

The above considerations show that the nonlinearity of the base resistance might become important at high frequencies and a high currents. However, when a transistor has to carry a

large base current, then the emitter area is usually larger than minimal, such that the effects that cause a change of the base resistance and, consequently, give rise to nonlinear behavior, are postponed to higher currents.

In Figure 8.17 the second harmonic and its different contributions are shown as a function of frequency. In correspondence with the circuit of Figure 8.16, the only capacitor that has been taken into account is  $C_{\pi}$ . The influence of other capacitors will be discussed in Section 8.2.6. The small-signal parameters and the nonlinearity coefficients that have been used in Figure 8.17 have been computed with the model parameters of Table 8.1. The transistor has been biased with a collector current of 0.8mA, which is one tenth of the knee current  $I_{KF}$ , and a base current of  $10\mu A$ . The cutoff frequency of the transistor, given by  $g_m/(2\pi C_{\pi})$ , is equal to  $13\,GHz$ .

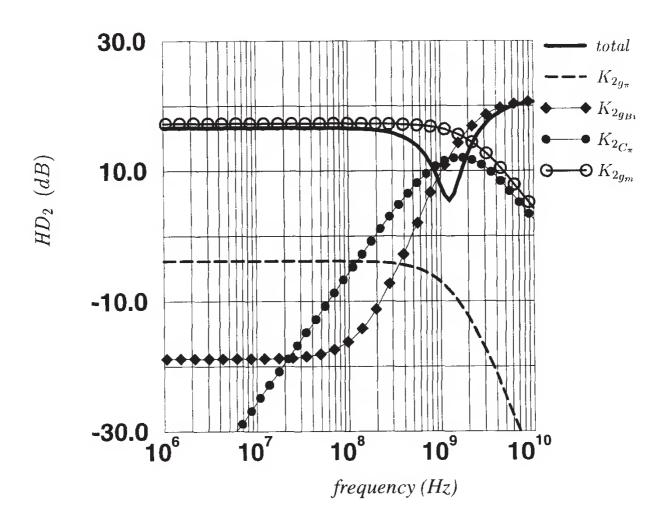


Figure 8.17: Second harmonic distortion of the output current of the single-transistor amplified of Figure 8.16 together with its different contributions, as a function of frequency. The source resistance has been made zero.  $HD_2$  is normalized to an input amplitude of 1V.

Figure 8.17 reveals that at low frequencies the collector current nonlinearity is the only significant second-order nonlinearity. At high frequencies, it is seen that the nonlinearity of the

base resistance becomes significant. It must be noted that the second-order nonlinearity coefficient of the base resistance has been computed with the power model (see Section 6.4), such that the contribution of this nonlinearity might be quite inaccurate. With a more accurate model, the contribution of the base resistance will be shifted downwards in the plot of Figure 8.17. In addition, we remark again that for high base currents the emitter area is usually larger than the minimal area that is allowed by the technology, such that the effects that cause a variation of the base current, are postponed to much higher currents.

The higher contribution of the base resistance nonlinearity to the second harmonic at high frequencies can be explained as follows: the input impedance of the transistor (without taking into account  $C_{\mu}$ ) is equal to the series connection of the base resistance with the parallel connection of  $r_{\pi}$  and  $C_{\pi}$ . At low frequencies and moderate bias current the input impedance is high since it is governed by  $r_{\pi}$ . At higher frequencies, the input impedance is lower:  $1/sC_{\pi}$  is now a much lower impedance than  $r_{\pi}$  and the input impedance is then given by the series connection of  $r_B$  and  $1/sC_{\pi}$ . Due to this lower input impedance, the AC current through the base resistance is higher. Also, the fraction of the input voltage over the base resistance is higher now compared to the low-frequency situation. Due to the higher swing of the voltage over the nonlinear base resistance, the second-order nonlinearity of the base resistance nonlinearity produces a relatively larger second-order signal.

In Figure 8.17 it is seen that at frequencies around  $1\,GHz$  the total second harmonic is smaller than the largest contributions. This can be explained by the differences in phase between the contributions. Whereas at low frequencies the different contributions are real numbers, each having a magnitude and a sign, the contributions at high frequencies are complex numbers, with a magnitude and a phase.

Analysis of the third harmonic distortion The third harmonic distortion and its contributions are shown as a function of frequency in Figure 8.18. The small-signal parameters and the secondand third-order nonlinearity coefficients are computed with the same model parameters as for the second harmonic distortion (see Table 8.1) and for the same bias currents ( $I_C = 0.8mA$  and  $I_B = 10\mu A$ ). In Figure 8.18 we have split the contribution of a nonlinearity into two parts, a first part proportional to the second-order nonlinearity coefficient of a nonlinearity and a second part that is proportional to the third-order nonlinearity coefficient. Such separation is possible since a third-order nonlinear current source falls apart in two such parts (see Table 5.7). The transfer function from the nonlinear third-order current source to the output of interest is of course the same for the two parts. In the rest of this chapter we will always split a third-order nonlinear current source in these two parts.

At low frequencies, the third harmonic is determined as discussed in Section 8.2.4.3 in which we considered the same nonlinearities except for  $C_{\pi}$ . We see that at the given collector current of 0.8mA the largest contributions to the third harmonic come from  $K_{2g_m}$  and  $K_{3g_m}$ .

At high frequencies, it is seen that the third harmonic is primarily determined by the second-order nonlinearities  $K_{2g_{Bi}}$ ,  $K_{2g_m}$  and  $K_{2C_{\pi}}$ . This is due to the large second-order signal at the controlling voltage of each of these nonlinearities. The controlling voltages are  $v_{BE}$  for  $K_{2g_m}$  and  $K_{2C_{\pi}}$ , and  $v_{B''B}$  for  $K_{2g_{Bi}}$ . At low frequencies, the signal swing over the intrinsic base resistance

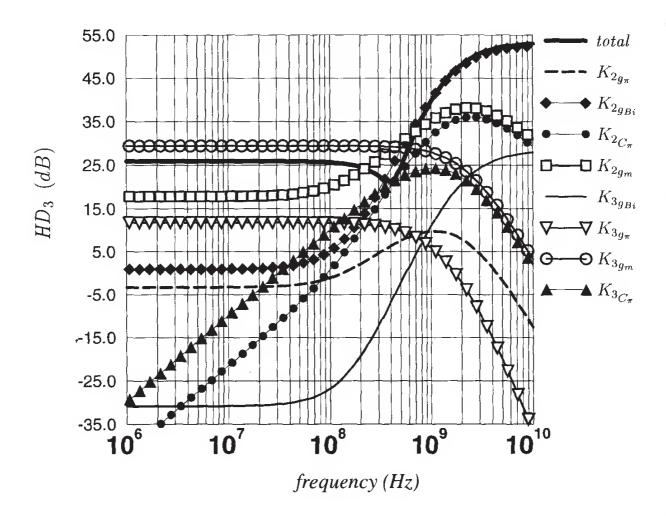


Figure 8.18: Third harmonic distortion of the output current of the single-transistor amplify of Figure 8.16 together with its different contributions, as a function of frequency. The source resistance has been made zero.  $HD_3$  has been normalized to an input amplitude of 1V.

is small and  $v_{be}$  is almost equal to the AC input voltage, which is a pure sinusoidal signal. higher frequencies, the fraction of the input voltage over  $r_{Bi}$  becomes larger. In this case, have a nonlinear voltage divider. The second-order signal over both  $r_{Bi}$  and the base-emit diode is significant. Each of the second-order nonlinearities  $K_{2g_{Bi}}$ ,  $K_{2g_m}$  and  $K_{2C_m}$  combinately such second-order signal with a first-order signal at its controlling terminals and produces third-order signal.

It must be noted once again that the influence of the base resistance nonlinearity at the given bias conditions is somewhat too high, due to the use of the power model, that yields overest mated values for the base resistance. Also, the presence of a nonzero source impedance, which is realistic at high frequencies, will lower the influence of the base resistance nonlinearity discussed in Section 5.5.1.

## **8.2.6** Influence of $C_{\mu}$ and $C_{cs}$

Finally, the transistor model is completed with the collector-substrate capacitor  $C_{cs}$  and the base-collector capacitance  $C_{\mu}$  of the transistor. If a load capacitor external to the transistor is present, it can simply be added to  $C_{cs}$ .

The collector-substrate junction and the base-collector are normally inversely biased. As a result, both  $C_{cs}$  and  $C_{\mu}$  only change slightly with the voltage over the junction such that the non-linearity coefficients of these capacitors are small. However, the value of the linearized capacitors  $C_{cs}$  and  $C_{\mu}$  will influence the nonlinear response as well.

Table 8.2 lists the small-signal parameters and the nonlinearity coefficients that are used in this section. They have been computed with the model parameters of Table 8.1. A difference with the previous sections is that here we will consider the base resistance as a linear element. Therefore, the model parameter  $I_{RB}$ , which is the current at which the intrinsic base resistance has fallen down to 50% of its value at very low base currents, has been made infinite.

With the inclusion of both  $C_{\pi}$  and  $C_{\mu}$  the transistor cutoff frequency  $f_T$  is now given by

$$f_T = \frac{g_m}{2\pi (C_\pi + C_\mu)} \tag{8.108}$$

With the small-signal parameters of Table 8.2 we find  $f_T = 12.2\,GHz$ .

The nonlinear circuit that is analyzed is shown in Figure 8.19. It contains five nonlinearities, namely three nonlinear capacitors, a nonlinear conductance and a two-dimensional nonlinear conductance. The output of interest is the collector-emitter voltage. The load resistance  $R_L$  is  $1k\Omega$ .

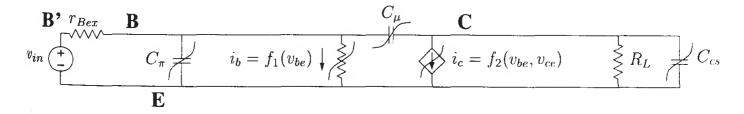


Figure 8.19: AC-equivalent circuit of the single-transistor amplifier of Figure 8.1 driven by a voltage source. The nonlinear capacitances  $C_{\pi}$ ,  $C_{\mu}$  and  $C_{es}$  and the nonlinear base and collector current are included.

## 8.2.6.1 Fundamental response

The linear response is computed with the linearized equivalent of the circuit of Figure 8.19. This linearized equivalent is shown in Figure 8.20.

First-order response of the output voltage The output of interest is the first-order component  $V_{ce,1,0}$  of the collector-emitter voltage. The ratio of  $V_{ce,1,0}$  and  $V_{in}$  is the voltage gain  $A_v$ . This

		<u> </u>	
$V_{BE}$	0.842V	$K_{2g_m}$	$0.459A/V^2$
$V_{CE}$	1.2V	$K_{3g_m}$	$4.00A/V^3$
$V_{CS}$	1.2V	$K_{2g_{\pi}}$	$7.02  mA/V^2$
$I_C$	0.834mA	$K_{3g_{\pi}}$	$91.0  mA/V^3$
$I_B$	$9.27 \mu A$	$K_{2g_o}$	$-3.10 \mu A/V^2$
$eta_{AC}$	90.0	$K_{3g_o}$	$1.40\mu A/V^3$
$g_m$	29.0mA/V	$K_{2_{g_m}\&g_o}$	$0.853  mA/V^2$
$r_{\pi}$	$2.77  k\Omega$	$K_{3_{2g_m}\&g_o}$	$13.5  mA/V^3$
$r_{Bi}$	93.0 Ω	$K_{3_{g_m}\&2g_o}$	$0.107  mA/V^3$
$r_{Bex}$	$120\Omega$	$K_{2_{C_{\pi}}}$	4.60~pF/V
$C_{\pi}$	0.353~pF	$K_{3}{}_{C_{\pi}}$	$40.1 \ pF/V^2$
$C_{\mu}$	24.4 fF	$K_{2_{C_{\mu}}}$	-3.81fF/V
$C_{cs}$	34.7fF	$K_{3_{C_{\mu}}}$	$1.60 fF/V^2$
$g_o$	$24.5\mu A/V$	$K_{2_{C_{cs}}}$	-3.21 fF/V
		$K_{3}_{C_{cs}}$	$0.793  fF/V^2$

Table 8.2: Bias conditions and the corresponding small-signal parameters and nonlinearity conficients for the bipolar transistor in the single-transistor amplifier of Figure 8.19.

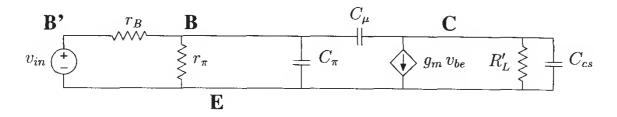


Figure 8.20: Linearized equivalent circuit of Figure 8.19.

result is

$$A_{v} = \frac{V_{ce,1,0}}{V_{in}} = -g_{m}R'_{L} \frac{1 - s\frac{C_{\mu}}{g_{m}}}{1 + \frac{r_{B}}{r_{\pi}} + \left[r_{B}\left(C_{\pi} + MC_{\mu}\right) + R'_{L}\left(1 + \frac{r_{B}}{r_{\pi}}\right)C_{cs}\right]s + s^{2}r_{B}R'_{L}C^{2}}$$
(8.109)

in which

$$R_L' = r_o \parallel R_L \tag{8.110}$$

$$M = 1 + \frac{R_L'}{r_B} + \frac{R_L'}{r_\pi} + g_m R_L'$$
 (8.111)

$$C^2 = C_{\pi}C_{\mu} + C_{\pi}C_{cs} + C_{\mu}C_{cs} \tag{8.112}$$

The expression for the gain can also be found in classical textbooks [Lak 94]. The factor M denotes the Miller factor [Lak 94].

The low-frequency gain as extracted from equation (8.109) is given by

$$A_{v0} = -g_m R_L' \frac{r_\pi}{r_\pi + r_R} \tag{8.113}$$

With the small-signal parameters of Table 8.2 and with  $R_L = 1k\Omega$  this yields a value of 26.3 or 28.4dB.

It is seen in equation (8.109) that  $A_v$  has two poles. A numerical computation of the two poles with the small-signal parameters of Table 8.2 yields values of 630MHz and 9.58GHz. Since these values are widely separated, one can approximate well the two poles by the ratio of coefficients of successive powers of s in the denominator of the transfer function [Lak 94]. This yields

$$p_1 \approx -\frac{\text{coefficient of } s^0}{\text{coefficient of } s^1}$$
 (8.114)

$$\approx -\frac{1}{\text{coefficient of } s^{1}}$$

$$= -\frac{1}{(C_{cs} + C_{\mu}) R'_{L} + (C_{\pi} + C_{\mu}) (r_{B} \parallel r_{\pi}) + g_{m} R'_{L} (r_{B} \parallel r_{\pi}) C_{\mu}}$$
(8.114)

Evaluation of this expression yields a pole at 591MHz, which is pretty close to the exact value of 630MHz. If the Miller effect would be stronger, either due to a larger  $C_{\mu}$ , a larger  $R'_{L}$  or a larger  $g_{m}$ , then expression (8.115) further reduces to

$$p_1 \approx \frac{1}{g_m R'_L(r_B \parallel r_\pi) C_\mu}$$
 (8.116)

However, with the current numerical values, this yields a poor approximation, since the Miller effect is small here.

The second pole can be approximated by

$$p_2 \approx -\frac{\text{coefficient of } s^1}{\text{coefficient of } s^2}$$
 (8.117)

$$= -\frac{r_B \left(C_{\pi} + MC_{\mu}\right) + R'_L \left(1 + \frac{r_B}{r_{\pi}}\right) C_{cs}}{r_B R'_L C^2}$$
(8.118)

which, for a strong Miller effect, reduces to

er effect, reduces to 
$$p_2 \approx -\frac{g_m C_\mu}{C_{cs} C_\pi + C_\mu C_{cs} + C_\mu C_\pi} \tag{8.119}$$

Finally, the zero in the expression of the voltage gain is found to be

$$z_1 = \frac{g_m}{C_u} {(8.120)}$$

For the current numerical values, this positive zero occurs at a frequency of 189 GHz.

For further computations it is interesting to formulate the first-order response in terms of the determinant of the admittance matrix. This determinant will be used in subsequent computations similarly to the computations in Section 5.3. When the admittance matrix formulation is used then conductances are used instead of resistances. In this way one finds

$$A_v = \frac{V_{ce,1,0}}{V_{in}} = \frac{g_B \left(-g_m + s C_\mu\right)}{\det(s)} \tag{8.121}$$

in which det(s) is the determinant of the admittance matrix, given by

$$\det(s) = (g_B + g_\pi)G'_L + s \left(C_\mu(g_B + g_\pi) + C_\mu g_m + C_{cs}(g_B + g_\pi) + C_\pi G'_L + C_\mu G'_L\right)$$

$$+ s^2 \left(C_{cs}C_\pi + C_{cs}C_\mu + C_\mu C_\pi\right)$$
(8.127)

The reader can verify that this is the same value as in equation (8.109).

**First-order response of**  $v_{BE}$  The first-order response  $V_{be,1,0}$  of the base-emitter voltage is important for further computations since this response determines the nonlinear current sources of order two and three of several nonlinearities. We will write  $V_{be,1,0}$  in terms of the determinant of the admittance matrix.

After some network analysis, either by hand or with ISAAC, we find, in terms of the determinant of the admittance matrix

$$V_{be,1,0} = \frac{g_B \left( G_L' + s \left( C_\mu + C_{cs} \right) \right)}{\det(s)} V_{in}$$
(8.123)

The poles in the expression of  $V_{be,1,0}$  have been discussed in the previous paragraph, The zero in the expression of  $V_{be,1,0}$  is given by

$$z_2 = -\frac{1}{R_L' \left( C_\mu + C_{cs} \right)} \tag{8.124}$$

In our numerical example, this zero occurs at a frequency of 2.75 GHz.

#### 8.2.6.2 Second harmonic distortion

The second harmonic distortion is composed of seven contributions. The five nonlinearities of Figure 8.19 give rise to seven nonlinear current sources: the two-dimensional collector current corresponds to three second-order nonlinearity coefficients, namely  $K_{2g_m}$ ,  $K_{2g_o}$  and  $K_{2g_m\&g_o}$ . These coefficients each give rise to a contribution to the second harmonic. The other four nonlinearities depicted in Figure 8.19 are one-dimensional, such that they each give rise to one contribution.

We will consider the second harmonic distortion not only of the output voltage, which is the collector-emitter voltage, but also of the base-emitter voltage. The reason is that the second-order component of the base-emitter voltage also determines the third harmonic distortion of the output voltage, as we will see in Section 8.2.6.3.

**Second harmonic distortion of the output voltage** Figure 8.21 depicts the second harmonic distortion of the output voltage of the single-transistor amplifier of Figure 8.19, together with its seven contributions as a function of frequency.

It is seen that at low frequencies there is one dominant contribution, namely the contribution of the coefficient  $K_{2g_m}$  that represents the dependence of the collector current on  $v_{BE}$ . The second harmonic distortion decreases at high frequencies. At frequencies above  $1\,GHz$  the contribution of  $K_{2G_m}$  becomes comparable to the contribution of  $K_{2g_m}$ . Since the two contributions partially cancel above  $1\,GHz$ , the total second harmonic distortion is smaller than these two contributions. As a result of this cancellation, the contribution of  $K_{2G_\mu}$  is significant as well in the frequency range from  $1.2\,GHz$  which is about one tenth of the cutoff frequency, up to a few GHz. Further it is seen that the contribution of the collector-substrate capacitor is very small. Between  $2\,GHz$  and  $3\,GHz$  it is larger than the total second harmonic distortion, due to the cancellation of larger contributions, but it remains one order of magnitude smaller than the largest contribution over the whole frequency band that is shown in Figure 8.21.

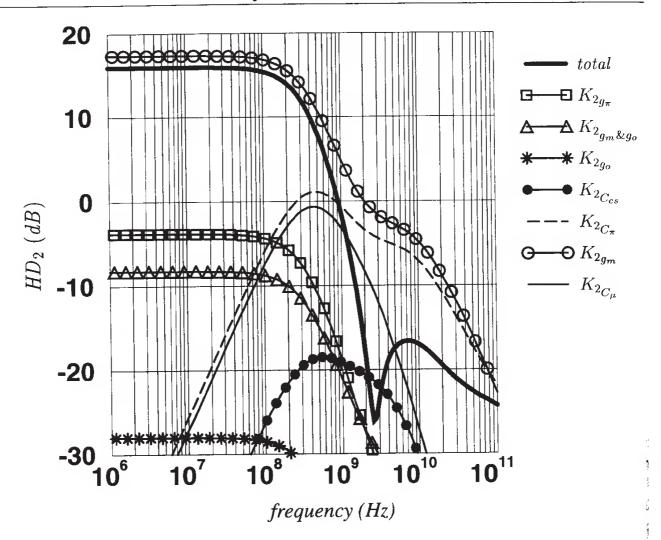


Figure 8.21: Second harmonic distortion at the output of the single-transistor amplifier of Figure 8.19, as a function of frequency together with its contributions.  $HD_2$  has been normalized to an input amplitude of 1V.

We will now determine a closed-form expression for the second harmonic distortion from low frequencies up to very high frequencies. We will take into account the two most important contributions, namely the contribution of  $K_{2g_m}$  and  $K_{2C_{\pi}}$ . If a higher accuracy is required then the contribution of  $K_{2C_n}$  should be taken into account as well.

With the two contributions to the second harmonic  $V_{out,2,0}$  of the output voltage, the second harmonic distortion of the output voltage is given by

$$HD_2 pprox rac{ ext{contribution of } K_{2g_m} ext{ to } V_{out,2,0} + ext{contribution of } K_{2C_\pi} ext{ to } V_{out,2,0}}{ ext{first-order response}}$$
 (8.125)

Using Table 5.5 we find that the contribution of  $K_{2g_m}$  to the second harmonic  $V_{out,2,0}$  is given by

contribution of 
$$K_{2g_m}$$
 to  $V_{out,2,0} = \frac{K_{2g_m}}{2} \cdot (V_{be,1,0})^2 \cdot TF_{i_{NL^2g_m} \to output}$  (8.126)

and for the contribution of  $K_{2C_{-}}$  we find

contribution of 
$$K_{2C_{\pi}}$$
 to  $V_{out,2,0} = sK_{2C_{\pi}} \cdot (V_{be,1,0})^2 \cdot TF_{i_{NL_2}C_{\pi}} \rightarrow output$  (8.127)

Hereby we remark that the transfer functions from the nonlinear current sources to the output must be evaluated for the frequency variable equal to 2s.

Summing the two contributions we find for the second harmonic  $V_{out,2,0}$ 

$$V_{out,2,0} = (V_{be,1,0})^2 \cdot \left(\frac{K_{2g_m}}{2} \cdot TF_{i_{NL2}g_m \to output} + sK_{2C_{\pi}} TF_{i_{NL2}C_{\pi} \to output}\right)$$
(8.128)

The transfer functions in this equation can be found with simple network analysis using the network of Figure 8.22. They are given by

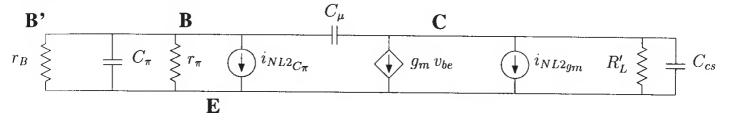


Figure 8.22: Linearized equivalent circuit of Figure 8.19 excited with the second-order nonlinear current sources  $i_{NL2_{gm}}$  and  $i_{NL2_{C\pi}}$ .

$$TF_{i_{NL2}g_m \to output} = \frac{1}{\det(2s)} \cdot (-g_B - g_\pi - 2s(C_\pi + C_\mu))$$
 (8.129)

and

$$TF_{i_{NL_2}C_{\pi}} \xrightarrow{output} = \frac{1}{\det(2s)} \cdot (-g_m + 2sC_{\mu})$$
(8.130)

Combining these transfer functions in equation (8.128) and using the value of  $V_{be,1,0}$  (equation (8.123)) we find

$$V_{out,2,0} = -\frac{g_B^2 \left(G_L' + s(C_\mu + C_{cs})\right)^2}{(det(s))^2 det(2s)} \cdot \left[\frac{K_{2g_m}}{2} \left(g_B + g_\pi + 2s(C_\pi + C_\mu)\right) + sK_{2C_\pi} \left(g_m - 2sC_\mu\right)\right] \cdot V_{in}^2$$
(8.131)

The second harmonic distortion is found by dividing  $V_{out,2,0}$  by  $V_{out,1,0}$ :

$$HD_{2} = \left| \frac{g_{B}^{2} (G_{L}' + s(C_{\mu} + C_{cs}))^{2}}{\det(s) \det(2s)(-g_{B}(g_{m} - sC_{\mu}))} \cdot \left[ \frac{K_{2g_{m}}}{2} (g_{B} + g_{\pi} + 2s(C_{\pi} + C_{\mu})) + sK_{2C_{\pi}}(g_{m} - 2sC_{\mu}) \right] \right| \cdot V_{in}$$
(8.132)

The denominator contains three factors: the determinant of the admittance matrix evaluated for s, the same determinant evaluated for 2s and the numerator of the first-order response  $V_{out,1,0}$ . This is in agreement with the general expression of a denominator of any second harmonic distortion, which is given in equation (5.132).

The low-frequency value of  $HD_2$  is found from equation (8.132):

$$HD_2$$
 (low frequencies) =  $\frac{K_{2g_m}}{2g_m} \cdot \frac{g_B}{g_B + g_\pi} \cdot V_{in}$  (8.133)

This result is in agreement with the result we found in equation (8.66).

We now consider the behavior of  $HD_2$  at higher frequencies. Since  $HD_2$  is a rational function in the frequency variable s we can analyze this in the same way as a transfer function of a linear circuit, namely using poles and zeros.

The following poles are found for  $HD_2$ :

- the poles of the first-order response. These are the zeros of det(s). In Section 8.2.6.1 these poles were found at 630MHz and 9.58GHz.
- poles at the half of the values of the poles of the first-order response. These are the zeros of det(2s). For the numerical example these occur at 315MHz and 4.79GHz.
- the zeros of the first-order response. This is due to the fact that  $HD_2$  is equal to the second-order response divided by the first-order response. From equation (8.121) we find that there is only one such zero in this circuit. This zero occurs at  $189\,GHz$ .

These poles are easily determined once the first-order response is known.

Next, we consider the zeros of  $HD_2$ . These are more difficult to determine, since the numerator of  $HD_2$  consists of a sum of numerators of contributions or, in other words, a sum of polynomials. We see from equation (8.132) that there is a double zero  $-G'_L/(C_{es}+C_\mu)$ . This occurs at  $2.75\,GHz$ . This double zero is the zero of  $V_{be,1,0}$  (see equation (8.123)). This zero occurs twice since  $V_{be,1,0}^2$  is a common factor of  $HD_2$  (see equation (8.128)). Finally, the second-order polynomial between the square brackets in the numerator of equation (8.132) gives rise to two zeros. These occur at  $600\,MHz$  and at  $218\,GHz$ . Since these zeros are so widely separated, a closed-form expression for these zeros can be found by taking ratios of the coefficients of this second-order polynomial, as we did in equations (8.115) and (8.118).

The poles and zeros that have been determined can now be used to check the frequency behavior of  $HD_2$ . The sharp minimum of  $HD_2$  around  $2.5\,GHz$  seen in Figure 8.21 cannot be explained with the above poles and zeros: this minimum occurs at a frequency where the largest contributions cancel, such that the contributions that were not taken into account play a role.

Second harmonic distortion of the base-emitter voltage Next, we determine the secondorder response at the base-emitter voltage. This response will be required to calculate the third harmonic distortion in the next section. The second harmonic at the base-emitter voltage is found by computing the base-emitter voltage in the linear circuit of Figure 8.20, which is excited with the nonlinear current sources of order two.

The second harmonic distortion of the base-emitter voltage and its most important contributions are shown as a function of frequency in Figure 8.23.

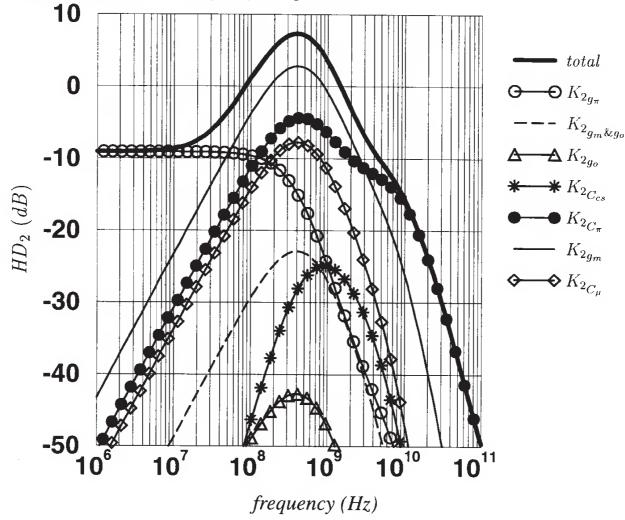


Figure 8.23: Second harmonic distortion of the base-emitter voltage in the single-transistor amplifier of Figure 8.19, as a function of frequency together with its most important contributions.  $HD_2$  has been normalized to an input amplitude of 1V.

It is seen that the second harmonic distortion at  $v_{BE}$  increases from about 10MHz, it reaches a maximum which is about 15dB higher than the low-frequency value of  $HD_2$  and then it decreases again. This "overshoot" is due to the contribution of the nonlinearity coefficient  $K_{2g_m}$ . This nonlinearity coefficient can only give a contribution to the second harmonic at  $v_{BE}$  at high frequencies: the nonlinear current source that corresponds to  $K_{2g_m}$  flows from the collector to the emitter. At low frequencies there is no path from the collector to the base-emitter voltage, such that the transfer function from the nonlinear current source to  $v_{BE}$  is zero. At high frequencies, however, there is such path along  $C_u$ .

We will now determine a closed-form expression for  $V_{be,2,0}$ . To this purpose, we will take into account the nonlinearities that give the most significant contributions to  $V_{be,2,0}$  from low frequencies up to  $100\,GHz$ . In this way, three second-order nonlinearities are seen to be significant: the base current nonlinearity, the nonlinearity of  $C_{\pi}$  and the nonlinearity of the collector current. With these three nonlinearities we will determine an approximate closed-form expression for the second harmonic distortion of  $v_{BE}$ .

The contribution of  $K_{2g_m}$  to the second harmonic  $V_{be,2,0}$  is found using Table 5.5. It is given by

contribution of 
$$K_{2g_m}$$
 to  $V_{be,2,0} = \frac{K_{2g_m}}{2} \cdot (V_{be,1,0})^2 \cdot TF_{i_{NL2}g_m \to v_{be}}$  (8.134)

in which the transfer function  $TF_{i_{NL2}g_m \to v_{be}}$  needs to be evaluated for the frequency variable 2s instead of s. Similarly, the contribution of  $K_{2g_{\pi}}$  is given by

contribution of 
$$K_{2g_{\pi}}$$
 to  $V_{be,2,0} = \frac{K_{2g_{\pi}}}{2} \cdot (V_{be,1,0})^2 \cdot TF_{i_{NL2}g_{\pi} \to v_{be}}$  (8.135)

and for the contribution of  $K_{2C_{\pi}}$  we find

contribution of 
$$K_{2_{C_{\pi}}}$$
 to  $V_{be,2,0} = sK_{2_{C_{\pi}}} \cdot (V_{be,1,0})^2 \cdot TF_{i_{NL_2_{C_{\pi}}} \to v_{be}}$  (8.136)

The transfer functions from the nonlinear current sources to the base-emitter voltage are found with network analysis using Figure 8.24. They are given by

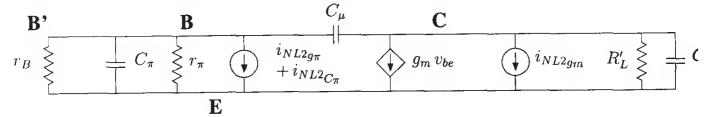


Figure 8.24: Linearized equivalent circuit of Figure 8.19 excited with second-order nonlinear current sources for the computation of  $V_{be,2,0}$ .

$$TF_{i_{NL^2gm} \to v_{be}} = -\frac{2s \, C_{\mu}}{\det(2s)}$$
 (8.137)

and

$$TF_{i_{NL2}g_{\pi} \to v_{be}} = TF_{i_{NL2}C_{\pi} \to v_{be}} = -\frac{G'_{L} + 2s\left(C_{cs} + C_{\mu}\right)}{\det(2s)}$$
 (8.138)

As already mentioned above, it is seen that the transfer function from the nonlinear second-order current source of  $K_{2g_m}$  to the base-emitter voltage is zero at low frequencies. It increases with frequency due to the path from the collector to the base along  $C_{\mu}$ .

Using the transfer functions in equations (8.134) to (8.136) we find for the second harmonic  $V_{be,2,0}$ :

$$V_{be,2,0} = \frac{V_{be,1,0}^2}{2\det(2s)} \cdot \left[ -K_{2g_m} 2s C_{\mu} - (K_{2g_{\pi}} + 2s K_{2C_{\pi}}) \cdot (G'_L + 2s (C_{cs} + C_{\mu})) \right]$$
(8.139)

The terms between the square brackets form a second-order polynomial. For use in the calculations of the third harmonic in the next section we rewrite  $V_{be,2,0}$  as follows:

$$V_{be,2,0} = \frac{V_{be,1,0}^2}{2\det(2s)} \cdot \left(a_0 + a_1s + a_2s^2\right) \tag{8.140}$$

with

$$a_0 = -K_{2q_{\pi}}G_L' \tag{8.141}$$

$$a_1 = -\left(2K_{2g_m}C_{\mu} + 2K_{2g_{\pi}}(C_{cs} + C_{\mu}) + 2K_{2C_{\pi}}G_L'\right)$$
(8.142)

$$\dot{a_2} = -4K_{2C_{\pi}}(C_{cs} + C_{\mu}) \tag{8.143}$$

From equation (8.140) it is seen that  $V_{be,2,0}$  has four poles: these are the poles of  $V_{be,1,0}$  and they occur twice, since  $V_{be,2,0}$  is proportional to the square of  $V_{be,1,0}$ . Using the same numerical data as in Section 8.2.6.1 we find a double pole at 630MHz and a double pole at  $9.58\,GHz$ . Further, we know from equation (8.123) that  $V_{be,1,0}$  has one zero, at  $2.75\,GHz$ . This zero is a double zero of  $V_{be,2,0}$ . Further, the zeros of the second-order polynomial  $(a_0 + a_1s + a_2s^2)$  are zeros of  $V_{be,2,0}$  as well. These zeros occur at 35.5MHz and at  $4.71\,GHz$ . Since these zeros are widely separated, a closed-form expression for these zeros can be found by taking ratios of coefficients, as we did in equations (8.115) and (8.118). Doing so, one finds for the low-frequency zero

$$z_1 \approx -\frac{a_0}{a_1} \approx \frac{K_{2g_{\pi}}G_L'}{2K_{2g_m}C_{\mu}}$$
 (8.144)

With the information about the poles and zeros the reader can easily verify the frequency behavior of  $V_{be,2,0}$ .

#### 8.2.6.3 Third harmonic distortion

The third harmonic distortion of the output voltage and its most significant contributions are shown in Figure 8.25. The contributions of  $K_{2_{C_{cs}}}$  and  $K_{3_{C_{cs}}}$  are insignificant in this example and they are not shown. Further, the contributions regarding the Early effect can be neglected as well. This means that the coefficients  $K_{2g_o}$ ,  $K_{3g_o}$ ,  $K_{2g_m\&g_o}$ ,  $K_{32g_m\&g_o}$  and  $K_{3g_m\&2g_o}$  are not important in this example.

It is seen that at low frequencies two contributions play a role, namely the contributions of  $K_{3g_m}$  and  $K_{2g_m}$ . They have an opposite sign, such that the total value of  $HD_3$  is smaller than the value of the largest contribution. This was already seen in Section 8.2.3.3, equation (8.74). From

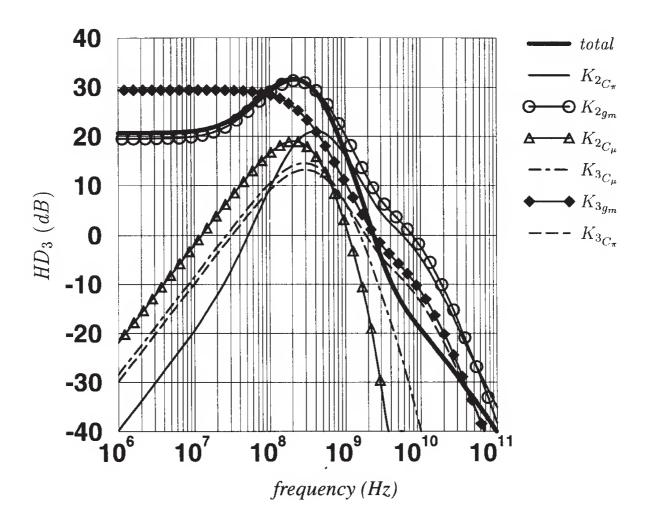


Figure 8.25: Third harmonic distortion at the output of the single-transistor amplifier of Figure 8.19, as a function of frequency together with its most important contributions.  $HD_3$  has been normalized to an input amplitude of 1V.

Table 5.7 we know that the contribution of  $K_{2g_m}$  is proportional to  $V_{be,2,0}$ . This explains why we analyzed  $V_{be,2,0}$  in the previous section.

The total value of  $HD_3$  begins to increase from about 10MHz. This can be explained as follows: at low frequencies the imaginary part of the contributions of  $K_{3g_m}$  and  $K_{2g_m}$  is zero. From the previous section we know that  $V_{be,2,0}$  has a zero at 35.5MHz. The contribution of  $K_{2g_m}$  also has this zero since it is proportional to  $V_{be,2,0}$ . As a result, the phase of this contribution goes to 90 degrees beyond 35.5MHz. On the other hand, the frequency behavior of the contribution of  $K_{3g_m}$  remains flat up to nearly 100MHz, which means that the phase is still about zero. As a result, the two contributions will not cancel anymore, and when their magnitude is equal, a about 80MHz, then the sum of the two contributions is not zero.

The third harmonic distortion reaches a maximum value around 200MHz. This maximum is about 10dB higher than the low-frequency value of  $HD_3$ . Beyond 200MHz,  $HD_3$  falls off.

We will now determine an approximate expression for the third harmonic distortion whice will allow us to explain the frequency behavior of  $HD_3$  by determining poles and zeros of  $HD_3$ 

In order to determine this expression, we will take into account the most important contributions. Apart from the ones from  $K_{3g_m}$  and  $K_{2g_m}$ , it is seen that at high frequencies the contribution of  $K_{3c_{\pi}}$  and  $K_{2c_{\pi}}$  play a role as well. Hence we will take into account four nonlinearity coefficients during the computations.

Using Table 5.7 the contribution of  $K_{2g_m}$  to the third harmonic  $V_{out,3,0}$  of the output voltage is given by

contribution of 
$$K_{2g_m}$$
 to  $V_{out,3,0} = K_{2g_m} \cdot V_{be,1,0} \cdot V_{be,2,0} \cdot TF_{i_{NL3}} \rightarrow output$  (8.145)

The contribution of  $K_{2C_z}$  is given by

contribution of 
$$K_{2_{C_{\pi}}}$$
 to  $V_{out,3,0} = K_{2_{C_{\pi}}} \cdot V_{be,1,0} \cdot V_{be,2,0} \cdot TF_{i_{NL3_{C_{\pi}}} \to output}$  (8.146)

For the contribution of  $K_{3q_m}$  we find

contribution of 
$$K_{3g_m}$$
 to  $V_{out,3,0} = \frac{K_{3g_m}}{4} \cdot V_{be,1,0}^3 \cdot TF_{i_{NL3}g_m \to output}$  (8.147)

and, finally, for the contribution of  $K_{^3C_\pi}$  we have

contribution of 
$$K_{3C_{\pi}}$$
 to  $V_{out,3,0} = \frac{3s K_{3C_{\pi}}}{4} \cdot V_{be,1,0}^{3} \cdot TF_{i_{NL3}} - \sigma_{output}}$  (8.148)

The third harmonic  $V_{out,3,0}$  is given by the sum of the four above contributions. From equation (8.140) we know that  $V_{be,2,0}$  is proportional to  $V_{be,1,0}^2$ . Hence the above contributions have a common factor  $V_{be,1,0}^3$ , and we find for the third harmonic

$$V_{out,3,0} = V_{be,1,0}^{3} \cdot \left[ \left( \frac{K_{2g_{m}} \left( a_{0} + a_{1}s + a_{2}s^{2} \right)}{2 \det(2s)} + \frac{K_{3g_{m}}}{4} \right) \cdot TF_{i_{NL3}g_{m} \to output} + 3s \left( \frac{K_{2C_{\pi}} \left( a_{0} + a_{1}s + a_{2}s^{2} \right)}{2 \det(2s)} + \frac{K_{3C_{\pi}}}{4} \right) \cdot TF_{i_{NL3}C_{\pi} \to output} \right]$$
(8.149)

The two transfer functions functions that occur in the expression of  $V_{out,3,0}$  do not have to be recomputed. In Section 8.2.6.2 we already computed the transfer functions from nonlinear current sources of order two to the output. The transfer functions from the third-order nonlinear current sources are found by replacing 2s with 3s. Doing so, we find from equations (8.129) and (8.130)

$$TF_{i_{NL3}g_m \to output} = \frac{1}{\det(3s)} \cdot (-g_B - g_\pi - 3s(C_\pi + C_\mu))$$
 (8.150)

and

$$TF_{i_{NL3}C_{\pi} \to output} = \frac{1}{\det(3s)} \cdot (-g_m + 3sC_{\mu})$$
 (8.151)

 $HD_3$  is found by dividing  $V_{out,3,0}$  by  $V_{out,1,0}$  given in equation (8.121). Using this equation together with equations (8.123), (8.149), (8.150) and (8.151) we find

$$HD_{3} = \left| \frac{g_{B}^{3} (G'_{L} + s(C_{cs} + C_{\mu}))^{3} V_{in}^{2}}{4(\det(s))^{2} \det(2s) \det(3s) g_{B}(-g_{m} + sC_{\mu})} \right| \cdot \left| \left[ (2K_{2g_{m}} (a_{0} + a_{1}s + a_{2}s^{2}) + K_{3g_{m}} \det(2s)) \cdot (-g_{B} - g_{\pi} - 3s (C_{\pi} + C_{\mu})) + 3s \left( 2K_{2C_{\pi}} (a_{0} + a_{1}s + a_{2}s^{2}) + K_{3C_{\pi}} \det(2s) \right) (-g_{m} + 3s C_{\mu}) \right] \right|$$
(8.152)

It is seen that the denominator in equation (8.152) is the same as the general expression of the denominator of any third harmonic distortion, which is given in equation (5.134).

The low-frequency value of the third harmonic distortion of the output voltage is found from equation (8.152):

$$HD_{3}(\text{low frequencies}) = \left| \frac{g_{B}^{2}V_{in}^{2} \left(2K_{2g_{m}}(-K_{2g_{\pi}}G_{L}') + K_{3g_{m}}G_{L}'(g_{B} + g_{\pi})\right)}{4G_{L}'g_{m}(g_{B} + g_{\pi})^{3}} \right|$$
(8.153)

This expression is consistent with the result obtained in Section 8.2.3 where we did not take into account any frequency dependence.

We now consider the poles and zeros of  $HD_3$  in order to explain the frequency behavior. The following poles are found for  $HD_3$ :

- the poles of the first-order response. These are the zeros of det(s). It is seen in equation (8.152) that these poles occur twice. In Section 8.2.6.1 these poles were found at 630MHz and 9.58GHz.
- poles at the half of the values of the poles of the first-order response. These are the zeros of det(2s). For the numerical example these occur at 315MHz and 4.79GHz.
- poles at one third of the values of the poles of the first-order response. These are the zeros of det(3s). For the numerical example these occur at 210MHz and 3.19GHz.
- the zeros of the first-order response. This is due to the fact that  $HD_2$  is equal to the second-order response divided by the first-order response. From equation (8.121) we find that there is only one such zero in this circuit. This zero occurs at  $189\,GHz$ .

The above poles are easily determined once the first-order response is known.

Next we consider the zeros of  $HD_3$ . Since the transfer functions  $TF_{i_{NL3}g_m \to output}$  and  $TF_{i_{NL3}C_{\pi} \to output}$  that occur in the expression of  $V_{out,3,0}$  (equation (8.149)) are first-order polynomials, the terms between the square brackets in equation (8.149) form a fourth-order polynomial in the frequency variable s. For our numerical example we find two real zeros at 35.7MHz and

 $4.73\,GHz$ , and two complex conjugate poles at  $113\,GHz$ . The zero at  $35.7\,MHz$  is almost identical to the zero at  $35.5\,MHz$  in the expression of  $V_{be,2,0}$ . An approximate expression of this zero has been given in equation (8.144). Further, the zero of  $V_{be,1,0}$  is a triple zero of  $HD_3$  since  $HD_3$  is proportional to the third power of  $V_{be,1,0}$ . In Section 8.2.6.1 it was found that this zero occurs at  $2.75\,GHz$ .

With these poles and zeros the maximum of  $V_{out,3,0}$  around 200MHz can be explained.

## 8.2.6.4 Third harmonic distortion with different values of $C_{\mu}$

In the single-transistor amplifier the base-collector capacitor  $C_{\mu}$  gives rise to the Miller effect [Lak 94]. In fact this capacitor constitutes a shunt-shunt feedback: this kind of feedback lowers both the input impedance and the output impedance. From Section 4.8 we know that feedback can lower the distortion. We will check this by changing the parameters of the loop gain of this shunt-shunt feedback via  $C_{\mu}$ .

First we analyze the loop gain qualitatively. The loop gain of the shunt-shunt feedback increases with an increase of  $g_m$ , and the load resistance  $R_L$ , as can be seen in the expression of the Miller factor, equation (8.111). This loop gain is low at low frequencies, since  $C_\mu$  is then a high impedance that does not couple the collector to the base. The loop gain increases with frequency, it then reaches a maximum value and at high frequencies it decreases. Further, it increases as  $C_\mu$  increases.

The loop gain of the shunt-shunt feedback can be computed by splitting the single-transistor amplifier into a basic amplifier and a feedback network, just as we did in Section 4.8.6. In this case, the feedback network consists of nothing else but  $C_{\mu}$ . When computing the loop gain, the loading of the feedback network must be taken into account in the basic amplifier [Gray 93].

The computation of the loop gain is left to the reader. We will only analyze the result. A schematic representation of the loop gain T as a function of frequency is shown in Figure 8.26. Its maximum value  $T_{max}$  is given by

$$T_{max} = \frac{g_m C_{\mu}}{q_L (C_{\pi} + C_{\mu}) + q_{\pi} C_{\mu}} \tag{8.154}$$

This maximum value is obtained from a frequency  $f_1$ , given by

$$f_1 = \frac{g_{\pi}G_L}{2\pi \left(G_L(C_{\pi} + C_{\mu}) + g_{\pi}C_{\mu}\right)}$$
(8.155)

and the loop gain decreases from a frequency  $f_2$ , given by

$$f_2 = \frac{G_L(C_\pi + C_\mu) + g_\pi C_\mu}{2\pi \left(C_\pi C_\mu + C_\mu^2\right)}$$
(8.156)

With this knowledge the third harmonic distortion has been computed with different values of  $C_{\mu}$  and  $R_{L}$ . For this experiment, the transistor output resistance has been omitted, since otherwise a very large  $R_{L}$  would be shunted by a smaller  $r_{o}$  such that the effect of a large effective load resistance cannot be examined properly. When changing  $C_{\mu}$ , the nonlinearity coefficients

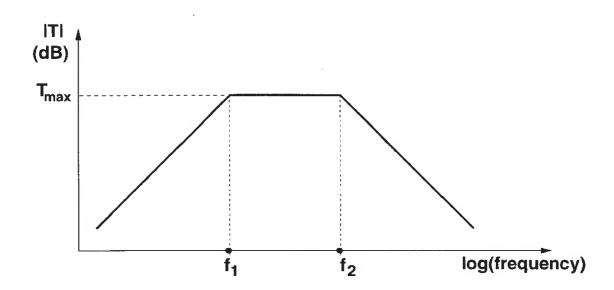


Figure 8.26: Bode diagram of the loop gain of the shunt-shunt feedback along  $C_{\mu}$  in a single-transistor amplifier.

 $K_{^2C_\mu}$  and  $K_{^3C_\mu}$  have been kept constant. The other small-signal parameters and nonlinearity coefficients have been adopted from the numerical example that has been used in the previous sections.

The third harmonic distortion as a function of frequency with different values for  $C_{\mu}$  and  $R_L$  is shown in Figure 8.27. The solid line in this figure corresponds to the same values as used in Figure 8.25, except that  $r_o$  has been omitted. For these values the maximum value of the loop gain, according to equation (8.154) is 1.8. When  $C_{\mu}$  is increased with a factor of ten and  $R_L$  is  $4k\Omega$ , then the maximum loop gain increases to 45. The frequencies  $f_1$  and  $f_2$ , according to equations (8.155) and (8.156), are found to be 60MHz and  $f_2=254MHz$ , respectively. For this case, shown with the dashed line in Figure 8.27, it is seen that the maximum of  $HD_3$  shifts to lower frequencies. Between 30MHz and 4GHz it is seen that  $HD_3$  is smaller than with the previous value of  $C_{\mu}$ .

When  $R_L$  is increased to  $10k\Omega$ , then from equation (8.154) a value of 104 is obtained for  $T_{max}$ , and  $f_1$  and  $f_2$  are equal to 38MHz and 159MHz, respectively. Compared to  $HD_3$  with  $R_L = 4k\Omega$ , it is seen that the maximum of  $HD_3$  is lower now.

For  $R_L = 100k\Omega$ ,  $T_{max}$  equals 488 and  $f_1$  and  $f_2$  are 5.95MHz and 101MHz, respectively. Despite the increase of the loop gain,  $HD_3$  is almost the same. This can be explained by the fact that the feedback network, in this case  $C_\mu$ , is nonlinear. In Section 4.8.2 it was proven that a large feedback can effectively suppress the nonlinearities in the basic amplifier but it cannot eliminate the effect of nonlinearities in the feedback network.

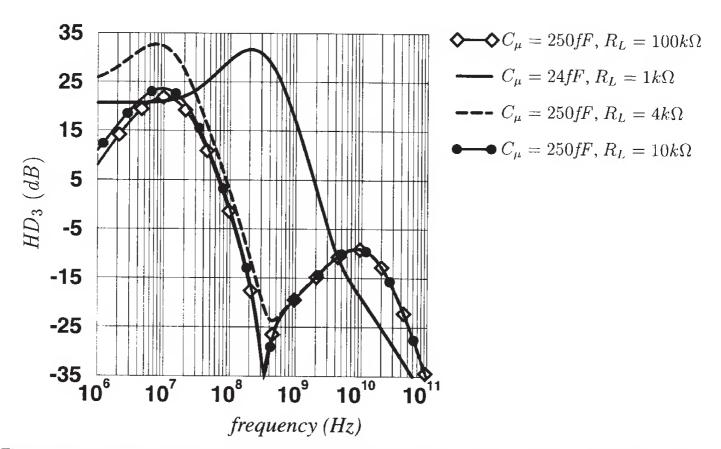


Figure 8.27: Third harmonic distortion at the output of the single-transistor amplifier of Figure 8.19 as a function of frequency for different values of  $R'_L$  and  $C_\mu$ .  $HD_3$  has been normalized to an input amplitude of 1V.

# 8.3 Single MOS transistor amplifier

Figure 8.28 shows an amplifier with one MOS transistor. The source and the bulk of the transistor are grounded. The input to the amplifier is a voltage source with an output resistance  $R_S$ . The amplifier is loaded with a resistance  $R_L$  and a capacitance  $C_L$ .

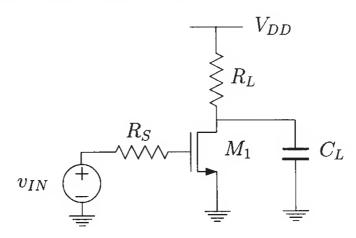


Figure 8.28: A single-transistor amplifier (MOS version).

We will analyze the harmonic distortion of this amplifier for different operating regions of the transistor and with the inclusion of different parasitics. First, the transistor will be represented by its (nonlinear) transconductance only. Next, the influence of the nonlinear output conductance will be modeled. Finally, the distortion at high frequencies will be considered for a MOS transistor in saturation. For the computations in this section many results can be taken over from the analysis of the bipolar single-transistor amplifier.

For numerical evaluations of the derived expressions we will make use of the technology data for the  $0.7\mu m$  process with the model parameters of Table 7.1.

## 8.3.1 Influence of $g_m$ only

We first compute the harmonics of the output voltage at low frequencies. This output voltage is the drain-source voltage of the transistor. In this section it is assumed that the drain current is a function of  $v_{GS}$  only. Since no output conductance is taken into account, and since the load resistance  $R_L$  is assumed to be linear, the distortion of the output voltage is the same as the distortion of the output current.

The starting point for the calculations is the nonlinear circuit of Figure 8.29 that is equivalent to Figure 8.28 for AC signals. In this circuit the nonlinear drain current is described a function of  $v_{gs}$ . This function is found from the model equation of the drain current. The circuit of Figure 8.29 is identical to its bipolar counterpart from Figure 8.2, apart from the functional dependence of the transistor current on the controlling voltage (gate-source or base-emitted voltage). The base current in Figure 8.2 does not play a role since  $R_S$  is zero. Hence, we calculate the results from Section 8.2.1.2 for the second and third harmonic distortion in terms of the

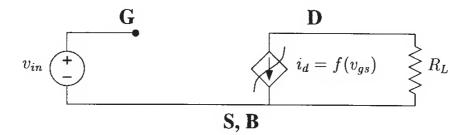


Figure 8.29: AC-equivalent circuit of the single-transistor amplifier of Figure 8.28 including the nonlinear dependence of the drain current on the gate-source voltage.

nonlinearity coefficients. The values of the nonlinearity coefficients, of course, will be different here.

With equation (8.36) we find that the second harmonic distortion of the output voltage or the output current is determined by the second-order normalized nonlinearity coefficient:

$$HD_2 = \left| \frac{K'_{2g_m}}{2} \right| V_{in} \tag{8.157}$$

and hence

$$IM_2 = \left| K'_{2q_m} \right| V_{in}$$
 (8.158)

For the third-order distortion we find from equation (8.40) that the third-order distortion is determined by the third-order normalized nonlinearity coefficient:

$$HD_3 = \left| \frac{K'_{3g_m}}{4} \right| V_{in}^2$$

$$IM_3 = \left| \frac{3K'_{3g_m}}{4} \right| V_{in}^2$$

The second-order intercept points  $IP_{2h}$  and  $IP_{3h}$  are then given by

$$IP_{2h} = \left| \frac{2}{K'_{2g_m}} \right|$$
 (8.162)

$$IP_{3h} = 2\sqrt{\left|\frac{1}{K'_{3g_m}}\right|}$$
 (8.163)

and for intermodulation distortion

$$IP_{2i} = \left| \frac{1}{K'_{2g_m}} \right| \tag{8.164}$$

$$IP_{3i} = 2\sqrt{\left|\frac{1}{3K_{3g_m}'}\right|} (8.165)$$

We will now evaluate these expressions in different operating regions.

### 8.3.1.1 Transistor in the saturation region

For the above distortion figures we will present expressions using different transistor models: we will use the level 1 model, since this is still widely used for hand calculations, as well as the model described in Section 7.7.2.4 that takes into account mobility reduction and velocity saturation.

Using the level 1 model If the MOS transistor in saturation is modeled with the level 1 model equation then  $K'_{2g_m}$  is given by equation (7.55), and we find

$$HD_2 = \frac{V_{in}}{4(V_{GS} - V_T)} \tag{8.166}$$

$$IM_2 = \frac{V_{in}}{2(V_{GS} - V_T)} \tag{8.167}$$

$$IP_{2h} = 4(V_{GS} - V_T) (8.168)$$

It is seen that the intercept point can be increased, or the linearity improved, by increasing  $(V_{GS} - V_T)$ . This is a design parameter that can be controlled. This situation is different from the single-transistor amplifier with a bipolar transistor. There it was seen that  $IP_{2h}$  only depends on the thermal voltage  $V_t$ .

The factor  $(V_{GS} - V_T)$  increases when  $V_{GS}$  increases, and it decreases when  $V_{SB}$  increases, since this increases the threshold voltage  $V_T$ . In this way, the bulk effect increases the nonlinear distortion.

Assume that the MOS transistor is biased such that  $(V_{GS} - V_T) = 200 mV$ . Then we find that  $IP_{2h}$  is 800 mV. This is about eight times larger than for a bipolar transistor.

Consider now the third-order distortion. Due to the quadratic dependence of the current on  $v_{GS}$  with the level 1 model, the third-order distortion figures  $HD_3$  and  $IM_3$  are zero and  $IP_{3h}$ ,  $IP_{3i}$  are infinite. This is an oversimplification, and in the next paragraph we will obtain more realistic values with a more accurate MOS model.

Using a more advanced model In Section 7.7.3 we used a drain current model that takes into account mobility reduction and velocity saturation. We will use the nonlinearity coefficients that

have been derived from this model in order to obtain closed-form expressions for the distortion figures.

An approximate closed-form expression for the normalized nonlinearity coefficient  $K'_{2g_m}$  has been derived in equation (7.139). Using this expression we find

$$HD_{2} = \frac{V_{in}}{2\left(1 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right)(V_{GS} - V_{T})\right) \cdot \left(2 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right)(V_{GS} - V_{T})\right) \cdot (V_{GS} - V_{T})}$$
(8.169)

It is seen that mobility reduction and velocity saturation lower the second harmonic distortion.

Setting  $HD_2$  to 1 and solving for  $V_{in}$  yields the second-order intercept  $IP_{2h}$  for harmonic distortion:

$$IP_{2h} = 2\left(1 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right) \cdot \left(2 + \left(\theta + \frac{\mu_0}{v_{sat}L}\right)(V_{GS} - V_T)\right) \cdot (V_{GS} - V_T)$$
(8.170)

Comparing this value to the intercept point computed with the level 1 model, it is seen that with the inclusion of mobility reduction and velocity saturation the intercept point is larger. This means that these effects cause a linearization of the transistor characteristics.

For a transistor with  $L=0.7\mu m$ ,  $V_{GS}=1.25V$  and  $V_{SB}=0V$ , and using the data of Table 7.1 we find  $V_{GS}-V_T=0.5V$ . Further,  $\theta=0.079V^{-1}$ , and  $\mu_0/v_{sat}/L=0.346V^{-1}$ . The factor  $(\theta+\mu_0/v_{sat}/L)\cdot(V_{GS}-V_T)$  equals 0.212. Then we find an intercept point  $IP_{2h}$  of 2.68V. Without mobility reduction and velocity saturation, equation (8.168) yields an intercept point of 2V.

Next, we consider the third harmonic distortion. An approximate expression for the third-order normalized nonlinearity coefficient  $K'_{3g_m}$  that takes into account mobility reduction and velocity saturation has been given in equation (7.141). Using this expression we find for  $HD_3$ 

$$HD_{3} = \frac{V_{in}^{2} \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right)}{4 \left(1 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right) (V_{GS} - V_{T})\right)^{2} \cdot \left(2 + \left(\theta + \frac{\mu_{0}}{v_{sat}L}\right) (V_{GS} - V_{T})\right) \cdot (V_{GS} - V_{T})}$$
(8.171)

It is seen that  $HD_3$  would be zero if mobility reduction and velocity saturation would not be present. This is not exact. Expression (7.141) for the third-order nonlinearity coefficient  $K'_{3g_m}$  that has been used to find the expression for  $HD_3$  is only an approximation. When mobility reduction and velocity saturation do not occur, then other terms in the exact expression for  $K'_{3g_m}$  become dominant. More precisely,  $K'_{3g_m}$  would be determined by the third-order derivative of the  $\frac{3}{2}$  powers that occur in the drain current due to the fact that the depletion layer underneath the channel varies along the channel.

Further it is seen that, just as for  $HD_2$ , the third harmonic distortion can be lowered by increasing  $V_{GS} - V_T$ .

By setting  $HD_3$  equal to one and solving for  $V_{in}$  one finds the third-order intercept point  $IP_{3h}$  for harmonic distortion. For the same transistor as in the numerical example for the computation of  $IP_{2h}$ , we find with  $L=0.7\mu m$  and  $V_{GS}-V_T=0.5V$ , a third-order intercept point  $IP_{3h}$  of 3.91V.

#### 8.3.1.2 Transistor in the triode region

Suppose that a MOS transistor in the triode region is used for amplification. This occurs for example in some integrators of active filters [Groen 94]. In this case, the nonlinearity coefficients can be found from equations (7.133) and (7.134). In this way we find for the second harmonic distortion of the output voltage or output current

$$HD_2 \approx \frac{1}{2} \frac{\theta}{1 + \theta f_u} \cdot V_{in} \tag{8.172}$$

and for the third-order distortion

$$HD_3 \approx \frac{1}{4} \frac{\theta^2}{(1 + \theta f_\mu)^2} \cdot V_{in}^2$$
 (8.173)

The coefficient  $f_{\mu}$  is given by equation (7.68), which is repeated here for convenience:

$$f_{\mu} = (v_{GB} - V_{FB} - \phi) - \frac{1}{2}(v_{DB} + v_{SB}) + \frac{2}{3} \frac{(\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2}}{v_{DB} - v_{SB}}$$
(8.174)

It is seen that  $HD_2$  and  $HD_3$  are only determined by mobility reduction. If no mobility reduction is present, then no distortion occurs. This is clear when one looks at the drain current expression in the triode region, equation (7.40), which shows a linear dependence of the drain current on  $v_{GS}$ .

Further, it is seen in equations (8.172) and (8.173) that the influence of velocity saturation is negligible.

As an example, we consider an n-MOS transistor in the  $0.7\mu m$  process with  $L=0.7\mu m$ ,  $W=16\mu m$ ,  $V_{GB}=3.2V$ ,  $V_{DB}=1.55V$  and  $V_{SB}=1.3V$ . Further  $\theta=0.05V^{-1}$ . Then we find  $f_{\mu}=3.717V$  and the second and third-order intercept points for harmonic distortion,  $IP_{2h}$  and  $IP_{3h}$  are both equal to about 47.2V, which is much larger than the intercept points for a MOS transistor in saturation.

## 8.3.1.3 Transistor in the weak inversion region

In the weak inversion region there is an exponential relationship between the drain current and the gate-source voltage. Then the expressions for the distortion figures closely resemble the expressions for the single-BJT amplifier with an elementary transistor model. We find for the

second- and third-order harmonic distortion of the output voltage or the output current:

$$HD_2 = \frac{V_{in}}{4nV_t} \tag{8.175}$$

$$HD_3 = \frac{V_{in}^2}{24n^2V_t^2} \tag{8.176}$$

in which n is the weak-inversion slope, defined in equation (7.208). It is given by

$$n \approx 1 + \frac{\gamma}{2\sqrt{1.5\Phi_F + v_{SB}}} \tag{8.177}$$

The second- and third-order intercept points for harmonic distortion,  $IP_{2h}$  and  $IP_{3h}$  are found by setting  $HD_2$  and  $HD_3$  to one:

$$IP_{2h} = 4nV_t \tag{8.178}$$

$$IP_{3h} = 2\sqrt{6}\,nV_t\tag{8.179}$$

In Section 7.13.1 we found a weak inversion slope of 1.275 for the  $0.5\mu m$  technology from Table 7.1. With this value,  $IP_{2h}$  is about 131mV and  $IP_{3h}$  is approximately 161mV at room temperature.

## 8.3.2 Influence of the output conductance

The simple transistor model used in the previous sections is now extended with an output conductance. As we saw in Section 7.11 this resistance is nonlinear. When the drain of the transistor is a high-impedance point, which means that  $R_L$  is high, then the effect of the output conductance and its nonlinearity can play a role in the harmonics of the current that flows into the external load resistance  $R_L$ . This has been discussed in Section 8.2.2 for a bipolar transistor. The results obtained there can be adopted and we will discuss them briefly here.

In several transistor models such as the BSIM3 model the drain current in saturation is described with a function of the form

$$i_{DSAT} = i_{DSAT}(v_{GS}, v_{SB}) \cdot g(v_{GS}, v_{SB}, v_{DS})$$
 (8.180)

in which  $g(v_{GS}, v_{SB}, v_{DS})$  models the output conductance, which is due to several phenomena, as described in Section 7.11.

If, on the other hand, the output conductance is assumed to be linear, then the drain current in the saturation regime is given by

$$i_{DSAT} = i_{DSAT} \left( 1 + \lambda v_{DS} \right) \tag{8.181}$$

For the computation of harmonics with the calculation method of Section 5.3, we start from the equivalent circuit of Figure 8.30. When the output resistance is included, then, both for a linear and nonlinear output resistance, the drain current is a function of two voltages. This function can be expanded into a two-dimensional power series. Compared to the previous section,

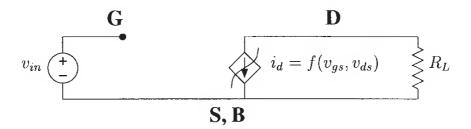


Figure 8.30: AC-equivalent circuit of the single-transistor amplifier of Figure 8.28 including the nonlinear dependence of the drain current on the gate-source and the drain-source voltage.

additional nonlinearity coefficients must be taken into account during the calculations, namely the coefficients that are proportional to second-order derivatives with respect to  $v_{DS}$  only and to both  $v_{GS}$  and  $v_{DS}$ .

The circuit of Figure 8.30 is identical to the circuit of Figure 8.4, apart from the functional dependence of the transistor current on the controlling voltages and apart from the base current. However, since the base current does not play a role in Figure 8.4, we can adopt the results from Section 8.2.2 in terms of the nonlinearity coefficients. The value of the coefficients, of course will be different here.

The second harmonic of the output current is given by

$$I_{out,2,0} = -\left(K_{2g_m} - K_{2g_m \& g_o} A_v + K_{2g_o} A_v^2\right) \frac{G_L}{G_L + g_o} \frac{V_{in}^2}{2}$$
(8.182)

in which  $A_v$  is the voltage gain  $g_m/(G_L+g_o)$ . It is seen that  $K_{2g_m\&g_o}$  contributes a factor  $A_v$  more to the second harmonic than  $K_{2g_m}$ . The coefficient  $K_{2g_o}$  contributes a factor  $A_v^2$  more to the second harmonic than  $K_{2g_m}$ .

Equation (8.182) reveals that when  $R_L$  is a large resistance, such that  $A_v$  is high and the drain of transistor  $M_1$  is a high-impedance node, then the contributions from  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  migh be considerable. This situation occurs for example when transistor  $M_1$  is at the output of the first stage of a two-stage amplifier with Miller compensation. On the other hand, when  $R_L$  is low, then the contribution of  $K_{2g_m}$  is dominant. In the limit, when  $R_L$  is a short circuit, then the contributions of  $K_{2g_m\&g_o}$  and  $K_{2g_o}$  vanish.

The result obtained in equation (8.182) can be intuitively explained as follows: when the voltage gain is high, then the voltage swing at the output is high. In this case, nonlinearities that are determined by the output voltage contribute much to the total harmonic.

Assume now that we consider the output conductance as a linear conductance. This mean that the coefficient  $K_{2g_o}$  is zero. On the other hand, the coefficient  $K_{2g_m\&g_o}$  is not zero. For example, for the level 1 model we find from Table 3.2

$$K_{2g_m \& g_o} = \mu_0 C'_{ox} \frac{W}{L} \lambda \ (V_{GS} - V_T)$$
 (8.183)

As a result, the second harmonic of the output current still contains a term that is related to the output conductance. This means that the presence of the output conductance contributes to the second harmonic, even when this conductance is assumed to be perfectly linear.

Next, we consider the third harmonic  $I_{out,3,0}$  of the output current. In accordance with equation (8.58) we find

$$I_{out,3,0} = -\frac{V_{in}^3}{4} \frac{G_L}{G_L + g_o} \left( K_{3g_m} + A_v^2 K_{3g_m \& 2g_o} - A_v^3 K_{3g_o} - A_v K_{32g_m \& g_o} \right.$$

$$+ \frac{A_v}{G_L + g_o} \left( K_{2g_m \& g_o} \right)^2 - \frac{K_{2g_m \& g_o} K_{2g_m}}{G_L + g_o} + \frac{2A_v^3}{G_L + g_o} K_{2g_o}^2$$

$$+ \frac{2A_v}{G_L + g_o} K_{2g_m} K_{2g_o} - 3 \frac{A_v^2}{G_L + g_o} K_{2g_m \& g_o} K_{2g_o} \right)$$

$$(8.184)$$

The terms on the first line of this equation arise from third-order nonlinearities (third-order nonlinearity coefficients) whereas the next terms are caused by second-order nonlinearities producing a third-order signal by combining a second-order signal with a first-order one. Similar conclusions can be drawn from equation (8.184) as from the expression of the second harmonic: when the drain is a low-impedance point, which means that  $R_L$  and  $A_v$  are low, then only  $K_{3g_m}$  gives a considerable contribution. In the case of a large value for  $R_L$  other contributions become important as well.

## 8.3.3 Frequency behavior

In this section we will analyze distortion at the output of the single MOS transistor amplifier of Figure 8.28 as a function of frequency. It is assumed that the transistor is in the saturation region. As a result, we can consider the capacitors  $C_{gs}$  and  $C_{gd}$  as linear capacitors (see Section 7.12). The output conductance will not be taken into account. It can always be added to the load conductance  $G_L = 1/R_L$ . Two nonlinearities will play a role in the single-transistor amplifier, namely the nonlinearity of the drain current and the nonlinearity of the drain-bulk capacitance. The other nonlinearities of the nonlinear equivalent circuit of a MOS transistor (Figure 7.1) are shorted such that they do not play a role. In this way, the single-transistor amplifier of Figure 8.28 reduces to the nonlinear equivalent circuit of Figure 8.31. The drain current nonlinearity is represented as a function  $f_1$  that relates the drain current to the gate-source voltage. The nonlinear drain-bulk capacitance is nothing else but a junction capacitance. It is described by a function  $f_2$  that relates the charge upon the capacitor to the drain-bulk voltage.

The small-signal parameters and the nonlinearity coefficients for the transistor that we will use in this section are listed in Table 8.3. The values have been obtained by computing the derivatives of the model equation (7.121). This equation does not take into account an output conductance. Further, the model parameters of the  $0.7\mu m$  process of Table 7.1 have been used, except that  $v_{sat}$  has been given the more realistic value of  $10^5 m/s$  and a value of  $0.05V^{-1}$  has been used for  $\theta$ , since a different mobility reduction model is used than the model for which the value of  $\theta$  from Table 7.1 has been fitted.

Further, the load resistance  $R_L$  is  $600\Omega$ . In this way, the low-frequency value of the gain  $v_{ds}/v_{in}$ , given by  $g_m R_L$ , equals 12dB. Further the source resistance is  $200\Omega$ . Finally, the capacitive load of the amplifier is taken equal to 100fF, which is about the same as the gate-source voltage of the transistor. We will analyze the circuit up to frequencies beyond the frequency

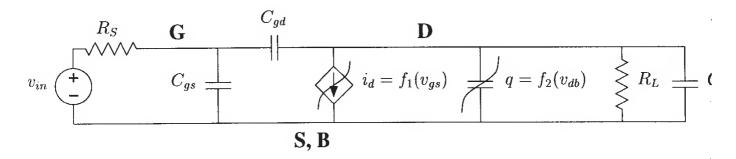


Figure 8.31: AC-equivalent circuit of the single-transistor amplifier of Figure 8.28. The nonlinear capacitance  $C_{db}$  and the nonlinear drain current are included.

W	80 μm	L	$0.7  \mu m$
$V_{GS}$	1.9V	$C_{gs}$	92.4fF
$V_{DS}$	2.25V	$C_{gd}$	16.8 fF
$V_{SB}$	0.0V	$C_{db}$	42.7 fF
$g_m$	6.61mA/V	$K_{2_{C_{db}}}$	-2.19 fF/V
$K_{2g_m}$	$1.81  mA/V^2$	$K_{3_{C_{db}}}$	$0.328 fF/V^2$
$K_{3g_m}$	$-0.592  mA/V^3$		

Table 8.3: Dimensions, bias conditions and the corresponding small-signal parameters and non linearity coefficients for the MOS transistor in the single-transistor amplifier.

range in which the nonlinear MOS model of Figure 7.1 is valid. This model is only valid for quesistatic operation, which is limited to a frequency  $0.1f_o$  with  $f_o$  given in equation (7.201). For the transistor here, we have an effective mobility  $\mu = 0.042m^2/(V.s)$ , the parameter a equal 1.2475, and  $(V_{GS} - V_T) = 1.15V$ , such that  $f_o = 12.6\,GHz$ . Nevertheless, an analysis up frequencies higher than  $0.1f_o$  can be useful to see some trends in the high-frequency behavior.

#### 8.3.3.1 First-order response

First we analyze the response of the linearized equivalent of Figure 8.31. This linearized circuits shown in Figure 8.32.

The output of interest is the drain-source voltage. Apart from this voltage we will also compute the first-order response of the gate-source voltage, since this determines the nonlinear computer than the contract of the gate-source voltage, since this determines the nonlinear contract of the gate-source voltage.

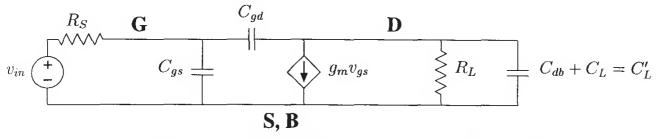


Figure 8.32: Linearized equivalent circuit of Figure 8.31.

rent sources of order two that correspond to the drain current nonlinearity.

The linearized circuit of Figure 8.32 has the same topology as the linearized equivalent circuit of the single-BJT amplifier from Figure 8.20 except that  $r_{\pi}$  is present there. Therefore we can adopt several results from Section 8.2.6.

First-order response of the output voltage The first-order component  $V_{ds,1,0}$  of the drain-source voltage can be computed by hand or with a symbolic network analysis program such as ISAAC. Just as with the computations with the bipolar transistor amplifier in Section 8.2.6 we will express the first-order response as a numerator that is divided by the determinant of the admittance matrix of the linearized network. This determinant will be used throughout the rest of the computations. Doing so we obtain

$$V_{out,1,0} = V_{ds,1,0} = \frac{G_S(-g_m + s C_{gd})}{\det(s)} \cdot V_{in}$$
(8.185)

in which det(s) is the determinant of the admittance matrix, given by

$$\det(s) = G_S G_L + s \left( C_{gd} G_S + C_{gd} g_m + C_L' G_S + C_{gs} G_L + C_{gd} G_L \right) + s^2 \left( C_L' C_{gs} + C_L' C_{gd} + C_{gd} C_{gs} \right)$$
(8.186)

in which  $C'_L = C_L + C_{db}$ .

We now consider the poles of the circuit. These are the zeros of the determinant det(s). Since this is a second-order polynomial, the circuit has two poles. When the Miller effect is strong, then the two poles of the circuit are widely separated. In this case, the two poles are given by

$$p_1 \approx \frac{G_S}{2\pi C_{gd} g_m R_L} \tag{8.187}$$

$$p_2 \approx \frac{g_m C_{gd}}{C_L' C_{gs} + C_{gd} C_L' + C_{gd} C_{gs}}$$
 (8.188)

These formulas can be found in several text books on analog integrated circuit design [Gray 93, Lak 94].

With the given numerical values, the Miller effect is quite small, due to the low gain of the amplifier. Then the approximations given in equations (8.187) and (8.188) are inaccurate. A better approximation is obtained by taking ratios of coefficients of different powers of s in det(s)as follows:

$$p_1 \approx - rac{ ext{coefficient of } s^0}{ ext{coefficient of } s^1}$$
 (8.189)
 $p_2 \approx - rac{ ext{coefficient of } s^1}{ ext{coefficient of } s^2}$  (8.190)

$$p_2 \approx -\frac{\text{coefficient of } s^1}{\text{coefficient of } s^2}$$
 (8.190)

Using these approximations we find poles at 1.2 GHz and at 10.2 GHz. An exact calculation yields a pole at 1.4 GHz and another one at 8.7 GHz. In reality, the high-frequency behavior will deviate from the behavior determined by the two poles, due to non-quasistatic behavior.

The expression for  $V_{ds,1,0}$  has one zero  $z_1$ :

$$z_1 = \frac{g_m}{2\pi C_{qd}} (8.191)$$

With the numerical values here, this zero occurs at  $63\,GHz$ .

First-order response of  $v_{GS}$  The first-order response  $V_{gs,1,0}$  of the gate-source voltage is found using network analysis. Similarly to the result from Section 8.2.6.1 for  $V_{be,1,0}$  we find

$$V_{gs,1,0} = \frac{G_S(G_L + s(C_{gd} + C_L'))}{\det(s)} V_{in}$$
(8.192)

The poles in the expression of  $V_{gs,1,0}$  have been discussed in the previous paragraph. The zero of  $V_{gs,1,0}$  is given by

$$z_2 = -\frac{1}{R_L(C_{qd} + C_L')} \tag{8.193}$$

In our numerical example this zero occurs at 1.67 GHz.

#### Second harmonic distortion 8.3.3.2

For the computation of the second harmonic distortion we apply to the linearized circuit two current sources of order two, as shown in Figure 8.33. One source corresponds to the nonlinearity of the drain current and the other one to the nonlinearity of the drain-bulk junction capacitor. These nonlinear current sources appear in parallel in the circuit. The contribution of each nonlinear current source is given by the product of its value and the transfer function from the nonlinear current source to the output of interest.

We will not only compute the second-order response of the output voltage. In addition, w will also compute the second-order response  $V_{gs,2,0}$  of the gate-source voltage. The reason that  $V_{gs,2,0}$  determines the nonlinear current source of order three that corresponds to the drain current nonlinearity. We will need the value of this current source in the computations of the third harmonic at the output.

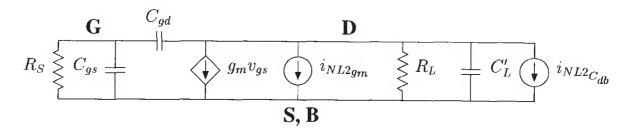


Figure 8.33: Linearized equivalent circuit of Figure 8.31 excited with the nonlinear current sources of order two.

Second harmonic distortion of the output voltage Figure 8.34 depicts the second harmonic distortion of the output voltage of the single-transistor amplifier of Figure 8.31, together with its contributions as a function of frequency.

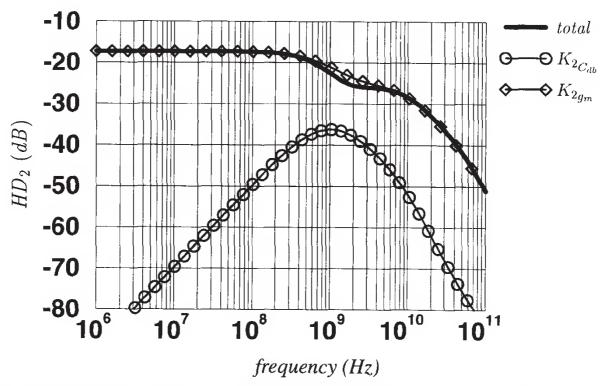


Figure 8.34: Second harmonic distortion at the output of the single-transistor amplifier of Figure 8.31, as a function of frequency together with its contributions.  $HD_2$  has been normalized to an input amplitude of 1V.

It is seen that there is one dominant contribution, namely the contribution of the coefficient  $K_{2g_m}$  that represents the nonlinear dependence of the collector current on  $v_{GS}$ . The contribution of  $C_{db}$  is small: at low frequencies it is small, since capacitive effects, both linear and nonlinear, do not play a role at low frequencies. At higher frequencies, the current through the capacitances  $C_L$  and  $C_{db}$  at the output increases. The part of the output current that flows through  $C_{db}$  yields additional distortion. One could expect that the distortion produced by  $C_{db}$  is high, since the

voltage swing over  $C_{db}$  is high. However,  $C_{db}$  is shunted by a larger linear capacitor  $C_L$ . Furthermore, the nonlinearity of  $C_{db}$  is small, since  $C_{db}$  is the capacitor of an inversely biased junction. The linearity of a junction capacitor improves when the junction is more reversely biased, as discussed in Section 3.4.

It is seen in Figure 8.34 that the second harmonic distortion decreases at high frequencies.

We will now determine a closed-form expression for the second harmonic distortion as a function of frequency. With the two contributions to the second harmonic  $V_{out,2,0}$  of the output voltage, the second harmonic distortion of the output voltage is given by

$$HD_2 pprox rac{ ext{contribution of } K_{2g_m} ext{ to } V_{out,2,0} + ext{contribution of } K_{2C_{db}} ext{ to } V_{out,2,0}}{ ext{first-order response}}$$
 (8.194)

Using Table 5.5 we find that the contribution of  $K_{2g_m}$  to the second harmonic  $V_{out,2,0}$  is given by

contribution of 
$$K_{2g_m}$$
 to  $V_{out,2,0} = \frac{K_{2g_m}}{2} \cdot (V_{gs,1,0})^2 \cdot TF_{i_{NL2}g_m \to output}$  (8.195)

and for the contribution of  $K_{2_{C_{ab}}}$  we find

contribution of 
$$K_{2C_{db}}$$
 to  $V_{out,2,0} = sK_{2C_{db}} \cdot (V_{ds,1,0})^2 \cdot TF_{i_{NL_2}C_{db}} \rightarrow output$  (8.196)

Hereby it must be noticed that the transfer functions from the nonlinear current sources to the output must be evaluated for the frequency variable equal to 2s. It is seen that the two nonlinear current sources are parallel to each other. Hence

$$TF_{i_{NL^{2}C_{db}} \to output} = TF_{i_{NL^{2}g_{m}} \to output}$$
(8.197)

This transfer function can be determined with network analysis. One finds

$$TF_{i_{NL^2g_m} \to output} = -\frac{G_S + 2s (C_{gs} + C_{gd})}{\det(2s)}$$
 (8.198)

If we only take into account the drain current nonlinearity — this is a reasonable assumption as seen from Figure 8.34 — then we find for  $V_{out,2,0}$  using equations (8.192) and (8.194) through (8.198)

$$V_{out,2,0} = -\frac{K_{2g_m}}{2} \cdot \frac{G_S^2 \left(G_L + s(C_L' + C_{gd})\right)^2 \left(G_S + 2s\left(C_{gs} + C_{gd}\right)\right)}{\left(\det(s)\right)^2 \det(2s)} \cdot V_{in}^2$$
(8.192)

The second harmonic distortion is found as the ratio of  $V_{out,2,0}$  and  $V_{out,1,0}$ . This yields

$$HD_{2} = \left| \frac{K_{2g_{m}}}{2} \cdot \frac{G_{S} \left( G_{L} + s \left( C'_{L} + C_{gd} \right) \right)^{2} \left( G_{S} + 2s \left( C_{gs} + C_{gd} \right) \right)}{\left( -g_{m} + s C_{gd} \right) \det(s) \det(2s)} \right| \cdot V_{in}$$
(8.20)

The frequency response of HD<sub>2</sub> can now be determined. At low frequencies we find that

$$HD_2$$
(low frequencies) =  $\frac{K_{2g_m}}{2g_m}V_{in}$  (8.201)

which corresponds to the result obtained in Section 8.3.1.

We now consider the behavior of  $HD_2$  at higher frequencies. Since  $HD_2$  is a rational function in the frequency variable s we can analyze this in the same way as a transfer function, namely using poles and zeros.

The following poles are found for  $HD_2$ :

- the poles of the first-order response. These are the zeros of det(s). In Section 8.3.3.1 these poles were found at  $1.4\,GHz$  and  $8.7\,GHz$ .
- poles at the half of the values of the poles of the first-order response. These are the zeros of det(2s). For the numerical example these occur at 700MHz and 4.35GHz.
- the zeros of the first-order response. This is due to the fact that  $HD_2$  is equal to the second-order response divided by the first-order response. From equation (8.185) we find that there is only one such zero in this circuit. This zero occurs at  $63\,GHz$ .

The above poles are easily determined once the first-order response is known.

Next, we consider the zeros of  $HD_2$ . The only contribution that we take into account in the expression of  $HD_2$  is the one from the nonlinearity of the drain current. This contribution is proportional to  $V_{gs,1,0}^2$ . Hence the zero of  $V_{gs,1,0}$  given in equation (8.193) is a double zero of  $HD_2$ . This double zero occurs at  $1.67\,GHz$ . Next, a zero occurs in the transfer function from the nonlinear current source to the output. From equation (8.198) we find that this zero occurs at a frequency  $1/(4\pi R_S(C_{gs}+C_{gd}))$ . For the numerical example, this frequency is  $3.65\,GHz$ . With these poles and zeros the frequency behavior of Figure 8.34 can be explained. If the contribution of  $K_{2C_{db}}$  is taken into account, then the numerator of  $HD_3$  becomes a third-order polynomial, yielding an additional zero.

**Second harmonic distortion at the gate-source voltage** Next, we determine the second-order response at the gate-source voltage. This response will be required to calculate the third harmonic distortion in the next section.

The second harmonic at the gate-source voltage is found by computing the gate-source voltage in the circuit of Figure 8.33, which is the linearized circuit excited with the nonlinear current sources of order two. The second harmonic distortion of the gate-source voltage and its contributions are shown as a function of frequency in Figure 8.35.

It is seen that at low frequencies there is no distortion at the gate-source voltage. This is explained as follows. The two sources of nonlinear distortion occur at the output. At low frequencies there is no path from the output (= the drain node) to the gate node. At higher frequencies, however, there is a path along  $C_{gd}$ . As a result, the second harmonic at the gate-source voltage increases with frequency. However, above  $1\,GHz$ , the second harmonic falls off with frequency.

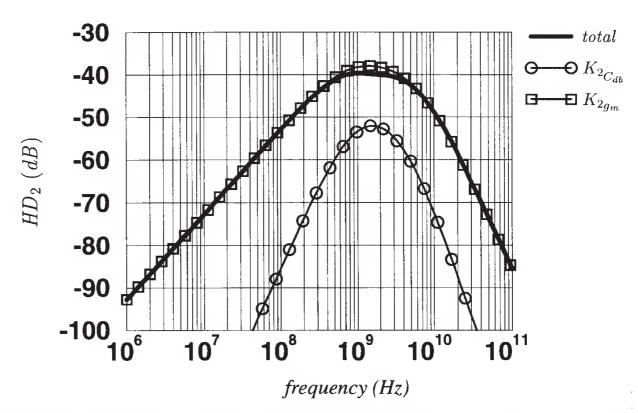


Figure 8.35: Second harmonic distortion at the gate-source voltage of the single-transistor amplifier of Figure 8.31, as a function of frequency together with its contributions.  $HD_2$  has been normalized to an input amplitude of 1V.

Further it is seen that along the complete frequency band the contribution of the second-order nonlinearity of the drain current is more than 10dB higher than the contribution of  $K_{2_{C_{db}}}$ . Hence an approximate expression for the second harmonic  $V_{gs,2,0}$  is given by

$$V_{gs,2,0} pprox rac{K_{2g_m}}{2} \cdot \left(V_{gs,1,0}
ight)^2 \cdot TF_{i_{NL2}g_m o v_{gs}}$$
 (8.202)

The transfer function from the nonlinear current source to the gate-source voltage is proportion to  $sC_{gd}$ , as we already mentioned. Indeed, with network analysis one finds

$$TF_{i_{NL2g_m} \to v_{gs}} = -\frac{2s C_{gd}}{\det(2s)}$$
(8.20)

Combining equations (8.192), (8.202) and (8.203) we find

$$V_{gs,2,0} = \frac{K_{2g_m}}{2} \frac{2sC_{gd}G_S^2 \left(G_L + s(C_{gd} + C_L')\right)^2}{\left(\det(s)\right)^2 \det(2s)} \cdot V_{in}$$
(8.20)

The frequency response of  $V_{gs,2,0}$  can now be analyzed. The poles of  $V_{be,2,0}$  are the same the poles of  $V_{out,2,0}$ . Further, the zero of  $V_{gs,1,0}$  (see equation (8.193)) at  $1.67\,GHz$  occurs twice With the knowledge of the first- and second-order responses at the controlling voltages, now have enough information to compute the third-order response.

#### 8.3.3.3 Third harmonic distortion

For the computation of the third harmonic, the nonlinear current sources of order three are applied to the linearized equivalent circuit of the single-transistor amplifier. The situation is completely similar to the computation of the second-order response with the nonlinear current sources of order two, as shown in Figure 8.33. The third harmonic distortion of the output voltage and its most significant contributions are shown in Figure 8.36.

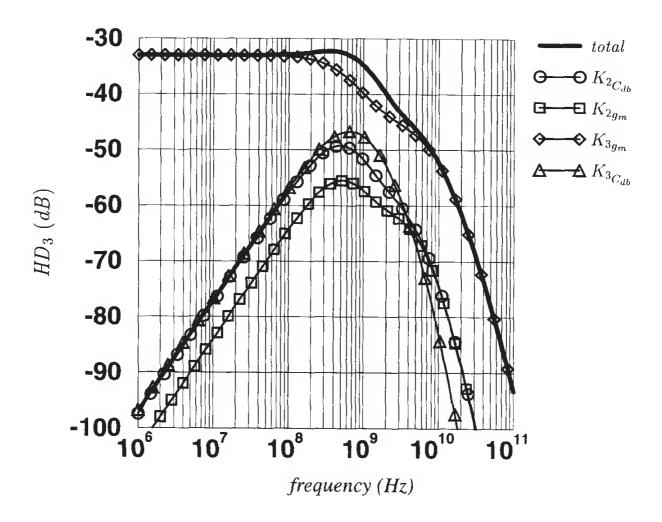


Figure 8.36: Third harmonic distortion at the output of the single-transistor amplifier of Figure 8.31, as a function of frequency together with its contributions.  $HD_3$  has been normalized to an input amplitude of 1V.

It is seen that at low frequencies there is only one significant contribution, namely the one of  $K_{3g_m}$ . If one would use the level 1 transistor model then this dominant contribution would be completely missing since with this model  $K_{3g_m}$  is zero. The magnitude of the contribution of  $K_{3g_m}$  starts to decrease from about 200MHz. Above 200MHz the total value of  $HD_3$  is higher than the contribution of  $K_{3g_m}$  since  $K_{3C_{db}}$  and the second-order nonlinearities give a contribution to the third harmonic distortion as well. The contributions of the second-order nonlinearity  $K_{2g_m}$  arises from the coupling of the drain node to the source node through  $C_{gd}$ . In this way, there is a

second-order component at the gate-source voltage. This component is combined by the second-order nonlinearity  $K_{2g_m}$  with the first-order component at the gate-source voltage, resulting in a third-order signal.

A closed-form expression for the third harmonic distortion as a function of frequency can be obtained by computing an expression for the different contributions and combining them. Here we will limit ourselves to the contribution of  $K_{3g_m}$  only, which is dominant over the entire frequency band. We find with Table 5.7

$$V_{out,3,0} \approx \frac{K_{3g_m}}{4} \cdot V_{gs,1,0}^3 \cdot TF_{i_{NL3}g_m \to out}$$
 (8.205)

The transfer function from the nonlinear current source to the output does not have to be recomputed. We can adopt it from the calculations of the second-order response. The only difference is that the frequency variable is now 3s instead of 2s. Using equation (8.198) we have

$$TF_{i_{NL3}g_m \to output} = -\frac{G_S + 3s\left(C_{gs} + C_{gd}\right)}{\det(3s)}$$
 (8.206)

Substituting this value into equation (8.205) and using the value of  $V_{gs,1,0}$  from equation (8.192) we find

$$V_{out,3,0} \approx -\frac{K_{3g_m}}{4} \cdot \frac{G_S^3 \left(G_L + s(C_L' + C_{gd})\right)^3 \left(G_S + 3s\left(C_{gs} + C_{gd}\right)\right)}{\left(\det(s)\right)^3 \det(3s)} \cdot V_{in}^3$$
(8.207)

Dividing  $V_{out,3,0}$  by  $V_{out,1,0}$  from equation (8.185) we obtain the third harmonic distortion:

$$HD_{3} \approx \frac{K_{3g_{m}}}{4} \cdot \left| \frac{G_{S}^{2} \left(G_{L} + s(C_{L}' + C_{gd})\right)^{3} \left(G_{S} + 3s\left(C_{gs} + C_{gd}\right)\right)}{\left(\det(s)\right)^{2} \det(3s)\left(-g_{m} + sC_{gd}\right)} \right| \cdot V_{in}^{2}$$
(8.208)

The low frequency value found from this equation is given by

$$HD_3$$
(low frequencies) =  $\frac{K_{3g_m}}{4g_m}V_{in}^2$  (8.209)

which corresponds to the result obtained in Section 8.3.1.

The frequency behavior is easily derived from the poles and zeros of the expression for  $HD_3$  is required that is more exact than the one from equation (8.208) the more contributions need to be taken into account.

# 8.4 Bipolar differential pair

Figure 8.37 shows a differential pair with bipolar transistors  $Q_{1A}$  and  $Q_{1B}$ . This circuit has a ready been analyzed qualitatively in Section 4.5.1. The differential pair is driven by a differential sinusoidal voltage  $v_{ID} = V_{id} \sin(\omega_1 t)$ . The output of interest is the voltage difference  $v_{OI}$  between the two collectors.

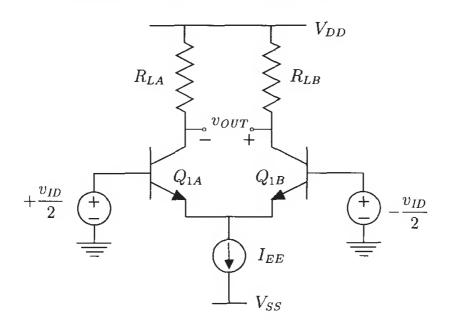


Figure 8.37: A bipolar differential pair.

We will limit the discussion here to the low-frequency nonlinear behavior. First, we will analyze the nonlinear distortion for an elementary transistor model. For this model it is easy to derive a closed-form expression for the DC transfer characteristic. Having such DC transfer characteristic, it can be developed into a power series. From the coefficients of this series it is possible to derive an expression for the harmonic and intermodulation distortion figures.

Next, we will compute distortion with the method of Section 5.3. This method is more general than using a DC transfer characteristic since this method allows to obtain closed-form expressions for the distortion even when a DC transfer characteristic cannot be computed.

A differential pair is a balanced circuit. When it is excited by a differential voltage and no mismatches are present in the circuit, then the even-order nonlinear responses are zero. This is only true if the components that have been designed identically, are really identical. In reality, such components have random differences in their behavior and they show a certain level of random mismatch in the parameters which model their behavior. This mismatch is due to the stochastic nature of physical processes that are used to fabricate the device. In Section 8.4.3 we will compute the second harmonic distortion in the presence of mismatches.

## 8.4.1 Computation of harmonics from the DC transfer characteristic

It is first assumed that the transistors  $Q_{1A}$  and  $Q_{1B}$  match, as well as the resistors  $R_{LA}$  and  $R_{LB}$ . If the transistors are modeled with their collector current only which satisfies the simple exponential relationship of equation (3.12), then a closed-form expression for the input-output relationship of a bipolar differential pair can be derived as explained in Section 4.5.1:

$$v_{OUT} = I_{EE}R_L \tanh\left(\frac{v_{ID}}{2V_t}\right) \tag{8.210}$$

in which  $R_L = R_{LA} = R_{LB}$ . Since capacitors have been neglected, the input-output relationship of equation (8.210) is also the DC transfer curve of the differential pair. This curve is depicted in Figure 4.15.

Having an explicit input-output relationship, the distortion figures can be easily derived. Using the power series expansion of a hyperbolic tangent around zero

$$\tanh(x) = x - \frac{x^3}{3} + \frac{2}{15}x^5 \dots \tag{8.211}$$

the input-output relationship can be approximated for sufficiently small signals as follows:

$$\tanh\left(\frac{v_{id}}{2V_t}\right) \approx \frac{v_{id}}{2V_t} - \frac{v_{id}^3}{24V_t^3} \tag{8.212}$$

This series can be identified with equation (2.4). Using equations (2.13) and (2.14) we find

$$HD_2 = 0$$
 (8.213)

$$HD_3 = \frac{V_{id}^2}{48V_t^2} \tag{8.214}$$

These distortion figures also apply to the differential output current, which is the difference of the current through the two load resistors: since the load resistors are linear and no output conductance of the transistors has been taken into account in the calculations, the ratio between the differential output voltage and the differential output current is just a factor  $R_L$ .

It is seen that the second harmonic distortion of the output voltage or current is zero. At analysis of the higher harmonics will show that all even-order harmonics are zero. This is general property of balanced circuits, as explained in Sections 2.2 and 4.5. The third harmonic is a factor two smaller than for the single-transistor of Section 8.2. If  $V_{id} = 25mV$  then  $HD_3 = 1/48 \approx 2.1\%$ .

The second- and third-order intercept points are derived from equations (8.213) and (8.214)

$$IP_{2h} = \infty ag{8.215}$$

$$IP_{3h} = 4\sqrt{3} \cdot V_t \tag{8.21}$$

At room temperature we find  $IP_{3h} \approx 0.179V$ . This is a factor  $\sqrt{2}$  higher than for the single-BJ amplifier.

At low frequencies and for small input amplitudes we know that  $IM_2 = 2HD_2$  and  $IM_3$   $3HD_3$  such that

$$IM_2 = 0$$
 (8.217)

$$IM_3 = \frac{V_{id}^2}{16V_t^2} \tag{8.21}$$

and

$$IP_{2i} = \infty ag{8.21}$$

$$IP_{3i} = 4V_t (8.22)$$

At room temperature we find  $IP_{3i} \approx 103 mV$ .

The higher-order even harmonics and intermodulation products of the output voltage or current are all zero when no mismatches are present. On the other hand, at the common-emitter point, only the even-order responses are not zero, whereas the odd-order responses are all zero, as explained in Section 4.5.1.

## 8.4.2 Symbolic analysis of $HD_3$

The method explained in Section 5.3 to compute harmonics, that has been implemented in the symbolic network analysis program ISAAC, is now used to compute the third harmonic at the output of the differential pair. The advantage of using this method is that closed-form expressions for the harmonics can be obtained also when no explicit input-output relationship such as equation (8.210) can be found.

The starting point for the calculations is the nonlinear circuit of Figure 8.38, that is equivalent with the differential pair of Figure 8.37. In this circuit, the base current and the collector current

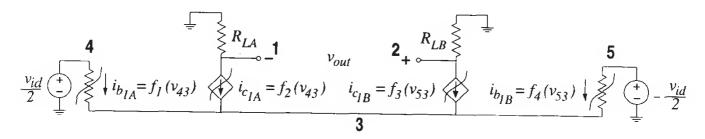


Figure 8.38: AC-equivalent circuit of the differential pair of Figure 8.37. Each transistor is modeled with a nonlinear collector current and a nonlinear base current.

of transistor  $Q_{1A}$  are modeled as a nonlinear function  $f_1$  and  $f_2$ , respectively, of the base-emitter voltage of  $Q_{1A}$ . Similarly, the functions  $f_3$  and  $f_4$  model the collector current and the base current of  $Q_{1B}$ . For the calculation of the third harmonic, the two transistors are assumed to match perfectly, just as the resistors. In this way, their small-signal parameters and nonlinearity coefficients are represented by the same symbol. This means for example that we will represent the transconductance of the two transistors with one symbol  $g_m$  rather than using two symbols. This will yield simpler expressions.

Further, the output resistance of the current source  $I_{EE}$  is assumed to be infinite.

The circuit of Figure 8.38 resembles the circuit of Figure 5.6, except that here the base currents are taken into account and only one excitation is applied. Interested readers can use the circuit of Figure 5.6 as a basis to compute the third harmonic by hand. Here we will use the symbolic network analysis program ISAAC that uses the calculation method of Section 5.3 to generate a closed-form expression.

First-order response First, the response of the linearized circuit to the differential voltage is computed. This circuit is shown in Figure 8.39.

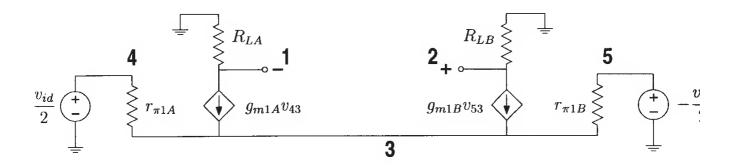


Figure 8.39: Linearized equivalent of the differential pair of Figure 8.38 used for the computation of the fundamental response.

Using ISAAC we find that the fundamental response of the output voltage is given by

$$V_{out,1,0} = V_{21,1,0} = \frac{2G_L g_m(g_m + g_\pi)}{\det(s)} \cdot V_{id}$$
(8.221)

with  $g_m = g_{m1A} = g_{m1B}$ ,  $G_L = G_{L1A} = G_{L1B}$  and  $g_{\pi} = g_{\pi 1A} = g_{\pi 1B}$ . The denominator o  $V_{out,1,0}$  is the determinant of the (C)MNA matrix, which has been found to be

$$\det(s) = 2G_L^2(g_m + g_\pi) \tag{8.222}$$

Since frequency dependence is not taken into account here, the determinant is equal for an frequency. Substituting equation (8.222) into equation (8.221) yields

$$V_{out,1,0} = \frac{g_m}{G_L} \cdot V_{id} \tag{8.223}$$

This is the well known result for a the response of a differential pair [Gray 93, Lak 94].

Next, we switch immediately to the third-order response. We know that the second-order response at the output is zero due to the balanced operation of the circuit. The second-order response at the base-emitter voltage of the two transistors, which will determine part of the third order nonlinear current sources, will not be computed explicitly. Its value will be directly take into account in the result of the third harmonic.

Third-order response With the calculation method of Section 5.3 the third-order response computed as the output of the linearized equivalent of Figure 8.38, excited with the appropriate current sources of order three. The value of these current sources can be determined with Table 5.7. This situation is shown in Figure 8.40.

It is seen that the current sources  $i_{NL3g_{\pi1A}}$  and  $i_{NL3g_{\pi1B}}$  are common-mode signals. When mismatches are present then they do not contribute to the third harmonic of the output voltage current. This means that the third harmonic will not depend on  $K_{3g_{\pi}}$  of one of the two transists.

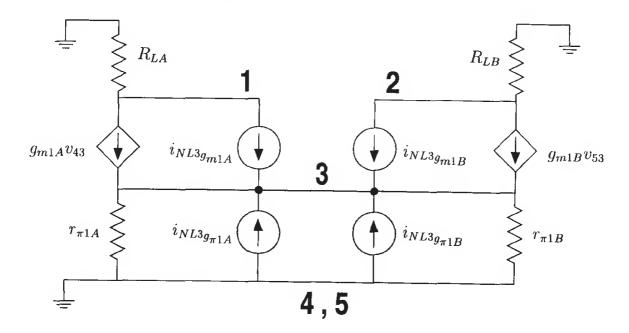


Figure 8.40: Linearized equivalent of the differential pair of Figure 8.38 used for the computation of the third harmonic of the differential output voltage.

Using ISAAC the following approximate expression has been computed for the third harmonic  $V_{out,3,0}$  of the output voltage

$$V_{out,3,0} = V_{21,3,0} \approx \frac{8G_L^9 \left(g_m + g_\pi\right)^4 \left(-K_{3g_m} \left(g_m + g_\pi\right) + 2K_{2g_m} \left(K_{2g_\pi} + K_{2g_m}\right)\right)}{4 \left(\det(s)\right)^3 \det(2s) \det(3s)} \cdot V_{id}^3$$
(8.224)

The only assumption that has been made to obtain this approximate expression is that  $g_m$  and its nonlinearity coefficients  $K_{2g_m}$  and  $K_{3g_m}$  are 80 times larger than  $g_\pi$ ,  $K_{2g_\pi}$  and  $K_{3g_\pi}$ , respectively. This corresponds to a transistor beta of 80. The error between the approximate result and the complete expression is less than 1%.

Substituting equation (8.222) into equation (8.224) yields, after some algebra

$$V_{out,3,0} = \frac{-K_{3g_m} (g_m + g_\pi) + 2K_{2g_m} (K_{2g_\pi} + K_{2g_m})}{16G_L(g_m + g_\pi)} \cdot V_{id}^3$$
(8.225)

The third harmonic distortion is found by dividing  $V_{out,3,0}$  by  $V_{out,1,0}$ . Using equations (8.223) and (8.225) we find

$$HD_{3} = \frac{-K_{3g_{m}} (g_{m} + g_{\pi}) + 2K_{2g_{m}} (K_{2g_{\pi}} + K_{2g_{m}})}{16(g_{m} + g_{\pi})g_{m}} \cdot V_{id}^{2}$$
(8.226)

f we put  $g_\pi$  and  $K_{2g_\pi}$  to zero, then we obtain

$$HD_3 = \frac{-K_{3g_m}g_m + 2K_{2g_m}K_{2g_m}}{16g_m^2} \cdot V_{id}^2$$
 (8.227)

Using the simple expressions for  $K_{2g_m}$  and  $K_{3g_m}$  given in equations (3.14) and (3.15), we find

$$HD_3 = \frac{V_{id}^2}{48V_t} \tag{8.228}$$

which corresponds to equation (8.214).

Although we obtained the same result seemingly with more effort than with the classical method of finding an expression for the DC transfer characteristic and developing it into a power series, the approach with ISAAC offers many advantages:

- The approach with ISAAC allows to take into account more nonlinearities than with the classical method. The latter does not work anymore if no explicit relationship between the input and the output can be found.
- The expressions are generic: they are function of the symbolic nonlinearity coefficients and small-signal elements, without actually using an explicit value for these symbolic parameters. This means that the expressions are valid for any model equation. The actual value of a parameter is only taken into account to approximate the symbolic expression.
- The expressions can be simplified with a user-defined error, such that the smaller terms do not disturb the interpretation of the dominant terms of an expression.
- As illustrated with other circuits in previous sections, it is possible to take into account frequency effects.

# 8.4.3 $HD_2$ due to mismatches

When there are mismatches between the transistors  $Q_{1A}$  and  $Q_{1B}$  or between  $R_{L1A}$  and  $R_{L1B}$ , then the circuit is no longer perfectly balanced. As a result, the even-order components of the differential output voltage or current are no longer zero. This can be seen as follows. Figure 8.41 shows the linearized circuit that is excited by the nonlinear current sources of order two for the computation of the second harmonic. The nonlinear current sources that are applied in this circuit are determined by the square of the base-emitter voltage of the corresponding transistor. When there are no mismatches in the circuit, then the base-emitter voltages of the two transistors have the same absolute value and an opposite sign. Hence the nonlinear current sources that correspond to transistor  $Q_{1A}$  are equal to the nonlinear current sources of transistor  $Q_{1B}$ . This means that the circuit of Figure 8.41 is perfectly symmetric, in the sense that no current flows across the dashed line. As a result, the voltages on nodes 1 and 2 are equal, such that the differential voltage in this circuit, which corresponds to the second harmonic, is zero.

When mismatches are present, then the circuit is not symmetric anymore, and a current flows across the dashed line in Figure 8.41. As a result, the voltages on nodes 1 and 2 will no longer be identical, such that a second harmonic occurs.

Due to mismatches the circuit parameters of corresponding devices such as the transconductance and the nonlinearity coefficients of transistors  $Q_{1A}$  and  $Q_{1B}$  are no longer identical. In

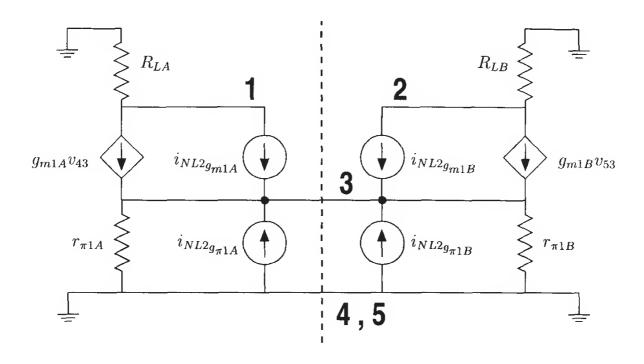


Figure 8.41: Linearized equivalent of the differential pair of Figure 8.38 used for the computation of the second harmonic of the differential output voltage.

order to describe mismatches, we define new parameters with the relations

$$\Delta X = X_A - X_B \tag{8.229}$$

$$X = \frac{X_A + X_B}{2} (8.230)$$

in which  $\Delta X$  is the difference between two parameters and X is the average of the two nominally matching parameters. Note that  $\Delta X$  can be positive or negative. In the differential pair we have for example

$$\Delta g_m = g_{m1A} - g_{m1B} \tag{8.231}$$

$$g_m = \frac{g_{m1A} + g_{m1B}}{2} \tag{8.232}$$

Equations (8.229) and (8.230) can be inverted:

$$X_A = X + \frac{\Delta X}{2}$$
$$X_B = X - \frac{\Delta X}{2}$$

For the actual computation of the second harmonic of the output voltage, we have assumed that there is a mismatch of 1% on the small-signal parameters and on the nonlinearity coefficients. This means for example that  $\Delta g_m$  is 1% of  $g_m$ .

The exact expression of the second harmonic is too complicated for an easy interpretation. Therefore, the exact expression has been approximated. The numerical values of the transistor parameters that have been used for this approximation are the ones from the single-transistor amplifier, listed in Table 8.2. These data correspond to a transistor beta of 80. Further the load resistance  $R_L$  is  $1k\Omega$  and  $\Delta R_L$  is  $10\Omega$ . With these numerical data ISAAC obtains the following approximate expression for the second harmonic  $V_{out,2,0}$  of the differential output voltage:

$$V_{out,2,0} = V_{21,2,0} = \frac{2G_L^5 g_m^2 \left(-3\Delta g_m K_{2g_m} + \Delta K_{2g_m} g_m\right)}{2\det(2s) \left(\det(s)\right)^2} \cdot V_{id}^2$$
(8.236)

This approximate expression consists of only two terms. The absolute value of the third largest term (that has been neglected) is a factor of about twenty five smaller than the sum of the absolute values of the two terms in the numerator of equation (8.236). This proves that the approximate expression is very accurate. Further, it is seen that the mismatches on  $g_{\pi}$  or  $K_{2g_{\pi}}$  are negligible. Also, the mismatch between  $R_{L1A}$  and  $R_{L1B}$  turn out to be negligible.

Using the expression for the determinant of the admittance matrix (equation (8.222)) and dividing  $V_{out,2,0}$  by  $V_{out,1,0}$  given in equation (8.223), the second harmonic distortion becomes

$$HD_2 = \frac{1}{8} \frac{g_m^3}{(g_m + g_\pi)^3} \left| -3 \frac{\Delta g_m}{g_m} K'_{2g_m} + \Delta K'_{2g_m} \right| \cdot V_{id}$$
 (8.237)

$$\approx \frac{1}{8} \left| -3 \frac{\Delta g_m}{q_m} K'_{2g_m} + \Delta K'_{2g_m} \right| \cdot V_{id}$$
(8.238)

Using the simple expression for  $K_{2q_m}$  of equation (3.14) yields

$$HD_2 = \frac{1}{8} \left| -\frac{3}{2V_t} \frac{\Delta g_m}{g_m} + \frac{\Delta g_m}{g_m} \frac{1}{2V_t} \right| \cdot V_{id} = \frac{1}{8} \frac{\Delta g_m}{g_m} \cdot \frac{V_{id}}{V_t}$$
 (8.239)

Hereby, we assumed that the mismatches on  $g_m$  and  $K_{2g_m}$  are correlated, such that they can be summed as normal numbers. If this would not be the case, then one should use standard deviations for the mismatch parameters  $\Delta g_m/g_m$  and  $\Delta K_{2g_m}$ .

The factor  $\Delta g_m/g_m$  in equation (8.239) can be related to a mismatch on the saturation current  $I_S$ , such that we obtain

$$HD_2 = \frac{1}{8} \frac{\Delta I_S}{I_S} \frac{V_{id}}{V_t}$$
 (8.240)

Mismatches in a differential pair give rise to an offset voltage. We will now relate the offset voltage to the second harmonic distortion computed in equation (8.239).

The offset voltage  $V_{OS}$  for a differential pair has been computed in [Gray 93]:

$$V_{OS} = V_t \left( \frac{\Delta R_L}{R_L} + \frac{\Delta I_S}{I_S} \right) \tag{8.241}$$

It is seen that the offset voltage has two contributions, one due to a mismatch on the load resistances, and the other one due to a mismatch on the saturation current  $I_S$  of the transistors  $Q_{1A}$  and  $Q_{1B}$ . Splitting the two terms in equation (8.241)

$$V_{OS} = V_{OS_{R_L}} + V_{OS_{I_S}} (8.242)$$

with

$$V_{OS_{R_L}} = V_t \frac{\Delta R_L}{R_L}$$
 and  $V_{OS_{I_S}} = V_t \frac{\Delta I_S}{I_S}$  (8.243)

we can substitute  $V_{OS_{I_S}}$  into the expression for  $HD_2$ :

$$HD_2 = \frac{1}{8} \frac{V_{OS_{I_S}}}{V_t^2} \cdot V_{id} \tag{8.244}$$

Hence we see that the second harmonic is proportional to the part of the offset voltage that is determined by mismatches on the saturation current.

The second-order intercept point  $IP_{2h}$  is determined by setting  $HD_2$  equal to one and solving for  $V_{id}$ . This yields

$$IP_{2h} = 8\frac{V_t^2}{V_{OS_{I_S}}} (8.245)$$

If  $\Delta I_S/I_S$  is 1% then we find that  $IP_{2h}$  is  $800V_t$ , which is 20.6V at room temperature. This is more than two orders of magnitude larger than the intercept point  $IP_{3h}$  for third-order harmonic distortion.

# 8.5 MOS differential pair

Figure 8.42 shows a differential pair with MOS transistors  $M_{1A}$  and  $M_{1B}$ . The differential pair is excited with a differential voltage  $v_{ID} = V_{id} \sin(\omega_1 t)$ .

For this circuit we will analyze the distortion of the differential output voltage  $v_{OUT}$  at low frequencies. We will follow the same approach as for the bipolar differential pair: first, the distortion figures will be derived from the DC transfer characteristic, using an elementary transistor model. Hereby the output conductance of the transistors will be neglected. As a result, the differential output voltage and the differential output current, which is the difference between the current through  $R_{LA}$  and  $R_{LB}$ , only differ by a constant factor  $R_L = R_{LA} = R_{LB}$ . In this way, the distortion on the output voltage is the same as on the output current.

Next, we will calculate a symbolic expression of the third harmonic. Here we will include the bulk effect. Finally, the second harmonic will be computed in the presence of mismatches.

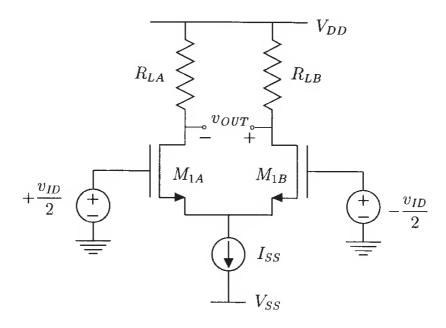


Figure 8.42: A MOS differential pair.

#### 8.5.1 Computation of harmonics from the DC transfer characteristic

In classical text books on analog integrated circuit design [Gray 93, Lak 94] a closed-form expression for the differential output voltage or output current is calculated using the level 1 MOS model. Hereby the transistors  $M_{1A}$  and  $M_{1B}$  are assumed to match perfectly and the output conductance of the transistors is neglected. With these assumptions one obtains for the differential output voltage

$$v_{OUT} = \frac{R_L I_{SS} v_{ID}}{V_{GS} - V_T} \sqrt{1 - \frac{v_{ID}^2}{4(V_{GS} - V_T)^2}}$$
(8.246)

in which  $V_{GS}$  is the quiescent value of the gate-source voltage of both  $M_{1A}$  and  $M_{1B}$ ,  $V_T$  is their threshold voltage, and  $R_L = R_{LA} = R_{LB}$ .

The DC transfer characteristic of equation (8.246) is valid for

$$|v_{ID}| < 2(V_{GS} - V_T) (8.247)$$

From the DC transfer characteristic of equation (8.246) the distortion figures are now derived. To this purpose, we develop the characteristic into a power series around  $V_{ID}=0$ . First, we introduce an auxiliary variable x given by

$$x = \frac{v_{id}}{2(V_{GS} - V_T)} \tag{8.248}$$

It can be remarked that we work here with the AC value  $v_{id}$ , whereas in equation (8.246) the total value  $v_{ID}$  is used. The use of  $v_{id}$  rather than  $v_{ID}$  is allowed since  $v_{ID} = v_{id} + V_{ID}$  and  $V_{ID} = 0$ .

With the new variable x from equation (8.248) the DC transfer characteristic around zero becomes

$$v_{out} = 2R_L I_{SS} x \sqrt{1 - x^2} (8.249)$$

The square root in this equation can be developed in a power series by making use of the relation

$$\sqrt{1-x^2} = 1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 + \dots \tag{8.250}$$

The power series expansion of the DC transfer characteristic of equation (8.249) now becomes

$$v_{out} = 2R_L I_{SS} \left( x - \frac{1}{2} x^3 - \dots \right)$$
 (8.251)

Substituting x by its value given in equation (8.248) we finally obtain

$$v_{out} = 2R_L I_{SS} \left( \frac{v_{id}}{2(V_{GS} - V_T)} - \frac{v_{id}^3}{16(V_{GS} - V_T)^3} - \dots \right)$$
 (8.252)

It is seen that there is no term in  $v_{id}^2$ . This is due to the balanced operation of the differential pair. Hence

$$HD_2 = IM_2 = 0 (8.253)$$

$$IP_{2h} = IP_{2i} = \infty ag{8.254}$$

The third harmonic distortion is found by making use of equation (2.14):

$$HD_3 = \frac{V_{id}^2}{32(V_{GS} - V_T)^2} \tag{8.255}$$

The third-order intercept point for harmonic distortion is found by setting  $HD_3$  equal to one and solving for  $V_{id}$ . This yields

$$IP_{3h} = 4\sqrt{2}\left(V_{GS} - V_T\right) \tag{8.256}$$

It is seen that  $IP_{2h}$  can be increased, and thus the linearity improved, by increasing  $(V_{GS} - V_T)$ . Hence, the design can influence the linearity of the differential pair. This is not the case for the bipolar equivalent (see equation (8.214)). For  $V_{GS} - V_T = 0.2V$  we find  $IP_{3h} = 1.13V$ . This is higher than the value of a bipolar differential pair, which is 179mV for any transistor bias condition in the forward active region and at room temperature.

At low frequencies and for small input amplitudes,  $IM_3$  is three times higher than  $HD_3$ :

$$IM_3 = \frac{3}{32} \cdot \frac{V_{id}^2}{(V_{GS} - V_T)^2} \tag{8.257}$$

and for the third-order intercept point  $IP_{3i}$  we find

$$IP_{3i} = 4\sqrt{\frac{2}{3}}\left(V_{GS} - V_T\right)$$
 (8.258)

For  $V_{GS} - V_T = 0.2V$  we find that  $IP_{3i} = 653 mV$ .

# 8.5.2 Computation of $HD_3$ including the bulk effect

We now compute an expression for  $HD_3$  in terms of the small-signal parameters and the nonlinearity coefficients by using the method of Section 5.3. Hereby, we will take into account the bulk effect. This means that the common source node is not connected to the bulk of the transistors  $M_{1A}$  and  $M_{1B}$ .

From classical text books on analog integrated circuit design [Gray 93, Lak 94] we know that the first-order response at the common source node is zero. However, we will find that the second-order response at this node is not zero. This second-order response will determine third-order nonlinear current sources that may give rise to significant contributions to the third-harmonic of the differential output voltage.

The starting point for the calculations is the nonlinear circuit of Figure 8.43, that is equivalent with the differential pair of Figure 8.42. In this circuit, the drain current of each transistor is

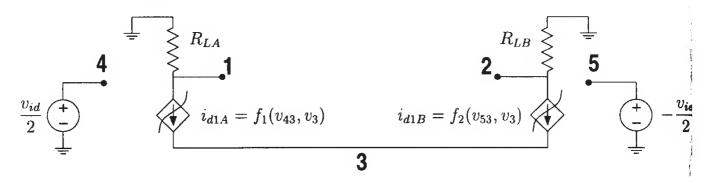


Figure 8.43: AC-equivalent circuit of the differential pair of Figure 8.42. Each transistor is modeled with a drain current that depends on the gate-source and the source-bulk voltage.

modeled as a function of the gate-source and the source-bulk voltage. The output conductance is neglected. For the calculation of the third harmonic it is assumed that there are no mismatches in the circuit. In this way, corresponding small-signal parameters and nonlinearity coefficients are represented by the same symbol in the final result. Further, the output resistance of the current source  $I_{SS}$  is assumed to be infinite.

**First-order responses** First, the response of the linearized circuit to the differential voltage is computed. This circuit is shown in Figure 8.44. The differential output voltage in this circuit is found to be

$$V_{out,1,0} = V_{21,1,0} = \frac{g_m}{G_L} \cdot V_{id}$$
 (8.259)

with  $g_m = g_{m1A} = g_{m1B}$  and  $G_L = G_{L1A} = G_{L1B}$ . The transconductance  $g_{mb} = g_{mb1A} = g_{mb1B}$  does not play a role here.

Further, we find that the first-order response at the common-source node is zero:

$$V_{3,1,0} = 0 (8.260)$$

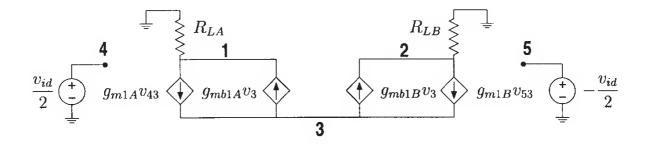


Figure 8.44: Linearized equivalent of the differential pair of Figure 8.43 used for the computation of the fundamental response.

Several nonlinear current sources of order two and three will be determined by the first-order response of the gate-source voltage of one of the two transistors. These are simply given by

$$V_{43,1,0} = \frac{V_{id}}{2} \tag{8.261}$$

$$V_{53,1,0} = -\frac{V_{id}}{2} \tag{8.262}$$

Second-order responses Next we compute the second-order responses. The linearized circuit that is excited with the nonlinear current sources of order two is shown in Figure 8.45. Since for

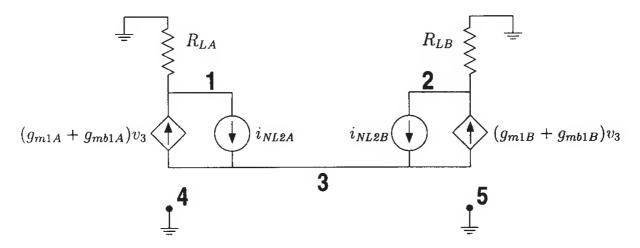


Figure 8.45: Linearized equivalent of the differential pair of Figure 8.43 excited with nonlinear current sources of order two for the computation of the second-order responses.

the computation of the second-order responses, the external voltage sources  $v_{id}/2$  and  $-v_{id}/2$  have to be shorted, the voltage-controlled current sources determined by  $g_m \cdot v_{gs}$  and  $g_{mb} \cdot v_{sb}$  are parallel for both transistors. Further, the nonlinear current source  $i_{NL2A}$  is the sum of the nonlinear current sources that correspond to the nonlinearity coefficients  $K_{2g_{m1A}}$ ,  $K_{2g_{mb1A}}$  and  $K_{2g_{m1A}\&g_{mb1A}}$ . These coefficients describe the second-order nonlinear dependence of the drain current on  $v_{GS}$  and  $v_{SB}$ . Using Table 5.5, we see that the current sources corresponding to the

coefficients  $K_{2g_{mb1A}}$  and  $K_{2g_{m1A}\&g_{mb1A}}$  are proportional to the first-order response  $V_{3,1,0}$  at the common-source node, which was found to be zero. As a result, the only component of  $i_{NL2A}$  that is not zero, is the one determined by  $K_{2g_{m1A}}$ :

$$i_{NL2A} = \frac{K_{2g_{m1A}}}{2} V_{43,1,0}^2 = \frac{K_{2g_{m1A}}}{8} \cdot V_{id}^2$$
 (8.263)

The nonlinear current source  $i_{NL2B}$  in Figure 8.44 is the sum of the nonlinear current sources that correspond to the nonlinearity coefficients  $K_{2g_{m1B}}$ ,  $K_{2g_{mb1B}}$  and  $K_{2g_{m1B}\&g_{mb1B}}$ . Again we find that only the current source that corresponds to  $K_{2g_{m1B}}$  differs from zero:

$$i_{NL2B} = \frac{K_{2g_{m1B}}}{2} V_{53,1,0}^2 = \frac{K_{2g_{m1B}}}{8} \cdot V_{id}^2$$
 (8.264)

It is seen that this nonlinear current source has the same value as  $i_{NL2B}$  when there are no mismatches in the circuit.

The differential output voltage in this network is the second-order response  $V_{out,2,0}$  at the output. With simple network analysis we easily find from Figure 8.45

$$V_{out,2,0} = V_{21,2,0} = \frac{i_{NL2A} - i_{NL2B}}{G_L}$$
 (8.265)

Since  $i_{NL2A} = i_{NL2B}$  we find the expected result

$$V_{out,2,0} = 0 (8.266)$$

For the second-order response at node 3 we find with simple network analysis

$$V_{3,2,0} = \frac{i_{NL2A} + i_{NL2B}}{2(g_m + g_{mb})} \tag{8.267}$$

Using the values of  $i_{NL2A}$  and  $i_{NL2B}$  from equations (8.263) and (8.264) we find

$$V_{3,2,0} = \frac{K_{2g_m}}{8(q_m + q_{mb})} \cdot V_{id}^2 \tag{8.268}$$

with  $K_{2g_m}=K_{2g_{m1A}}=K_{2g_{m1B}}$ . This expression is the second-order response of the source-bulk voltage of each of the two transistors  $M_{1A}$  and  $M_{1B}$ . We will use this expression in the calculation of the third harmonic. Further, we will need the second-order response of the gate-source voltage of the two transistors. The gates of the two transistors are fixed to the input voltage source, which is assumed to be a pure sinusoidal signal. Hence there are no higher-order signals present at the gate of  $M_{1A}$  and  $M_{1B}$ . As a result we find

$$V_{43,2,0} = V_{4,2,0} - V_{3,2,0} = 0 - V_{3,2,0} = -\frac{K_{2g_m}}{8(g_m + g_{mb})} \cdot V_{id}^2$$
(8.269)

and similarly

$$V_{53,2,0} = -\frac{K_{2g_m}}{8(g_m + g_{mb})} \cdot V_{id}^2 \tag{8.270}$$

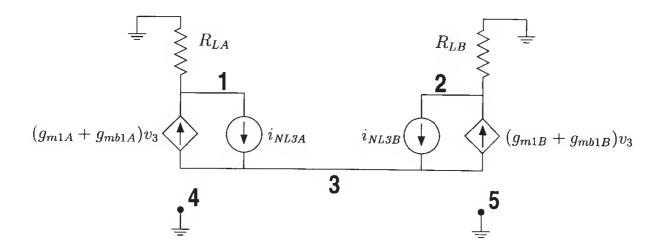


Figure 8.46: Linearized equivalent of the differential pair of Figure 8.43 excited with nonlinear current sources of order three for the computation of the third-order response.

**Third-order response** For the calculation of the third harmonic, the linearized circuit is excited with nonlinear current sources of order three as shown in Figure 8.46.

The third-order nonlinear current source  $i_{NL3A}$  shown in this figure, is the sum of the third-order nonlinear current sources that correspond to the nonlinearity coefficients  $K_{2g_{m1A}}$ ,  $K_{3g_{m1A}}$ ,  $K_{2g_{m1A}\&g_{mb1A}}$ ,  $K_{3g_{m1A}\&g_{mb1A}}$  and  $K_{3g_{m1A}\&2g_{mb1A}}$ . From Table 5.7 we see that the nonlinear current sources that correspond to the coefficients  $K_{3g_{m1A}\&2g_{mb1A}}$ ,  $K_{3g_{m1A}\&g_{mb1A}}$  are proportional to the first-order response  $V_{3,1,0}$  of the source-bulk voltage. Hence these sources are zero since we found that  $V_{3,1,0}=0$ .

From Table 5.7 we find that the third-order nonlinear current source that corresponds to the coefficient  $K_{2g_{m1A}\&g_{mb1A}}$ , consists of two parts:

3rd-order current source proportional to 
$$K_{2_{g_{m1A}\&g_{mb1A}}}=\frac{1}{2}K_{2_{g_{m1A}\&g_{mb1A}}}V_{43,2,0}V_{3,1,0} + \frac{1}{2}K_{2_{g_{m1A}\&g_{mb1A}}}V_{43,1,0}V_{3,2,0}$$
 (8.271)

The first term of this component is proportional to the first-order response  $V_{3,1,0}$  at the common source. Hence it is zero. The second term is not zero since  $V_{3,2,0}$  differs from zero.

We can conclude that the nonlinear current source  $i_{NL3A}$  consists of three parts that are not zero: a first part that is proportional to  $K_{2g_{m1A}\&g_{mb1A}}$ , a second part proportional to  $K_{2g_{m1A}}$  and a third part proportional to  $K_{3g_{m1A}}$ . Using Table 5.7 we find

$$i_{NL3A} = \frac{K_{3g_{m1A}}}{4} V_{43,1,0}^3 + K_{2g_{m1A}} V_{43,2,0} V_{43,1,0} + \frac{K_{2g_{m1A} \& g_{mb1A}}}{2} V_{43,1,0} V_{3,2,0}$$
(8.272)

Using equations (8.260), (8.261), (8.268) and (8.269) we find

$$i_{NL3A} = \frac{K_{3g_{m1A}}}{32} V_{id}^3 - \frac{K_{2g_{m1A}}}{2} V_{3,2,0} V_{id} + \frac{K_{2g_{m1A} \& g_{mb1A}}}{4} V_{3,2,0} V_{id}$$
(8.273)

For the third-order nonlinear current source  $i_{NL3B}$  we find that it is opposite to  $i_{NL3A}$  when there are no mismatches in the circuit. Hence

$$i_{NL3B} = -\frac{K_{3g_{m1B}}}{32}V_{id}^3 + \frac{K_{2g_{m1B}}}{2}V_{3,2,0}V_{id} - \frac{K_{2g_{m1B}\&g_{mb1B}}}{4}V_{3,2,0}V_{id}$$
(8.274)

The third harmonic of the output voltage  $V_{out,3,0}$  is equal to the differential output voltage in Figure 8.46. We easily find

$$V_{out,3,0} = V_{21,3,0} = \frac{i_{NL3A} - i_{NL3B}}{G_L}$$
 (8.275)

Since  $i_{NL3A} = -i_{NL3B}$  we have

$$V_{out,3,0} = V_{21,3,0} = \frac{2i_{NL3A}}{G_L} \tag{8.276}$$

Since we have assumed that there are no mismatches in the circuit, we now omit the subscripts in the small-signal elements and the nonlinearity coefficients that make a difference between the two transistors  $M_{1A}$  and  $M_{1B}$ . Using equations (8.268) and (8.273) we find for the third harmonic at the output

$$V_{out,3,0} = \frac{2}{G_L} \cdot \left( \frac{K_{3g_m}}{32} - \frac{K_{2g_m}^2}{16(g_m + g_{mb})} + \frac{K_{2g_m \& g_{mb}} K_{2g_m}}{32(g_m + g_{mb})} \right) \cdot V_{id}^3$$
(8.277)

The third harmonic distortion is found by dividing  $V_{out,3,0}$  by  $V_{out,1,0}$ . This yields

$$HD_3 = \frac{1}{8g_m} \cdot \left( \frac{K_{3g_m}}{2} - \frac{K_{2g_m}^2}{(g_m + g_{mb})} + \frac{K_{2g_m \& g_{mb}} K_{2g_m}}{2(g_m + g_{mb})} \right) \cdot V_{id}^2$$
 (8.278)

We first evaluate  $HD_3$  using the level 1 model for the transistor. Since this is a quadratic model, the nonlinearity coefficient  $K_{2q_m}$  is zero. Next, we find, using Table 3.2,

$$\frac{K_{2g_m}}{g_m} = -\frac{K_{2g_m \& g_{mb}}}{2g_{mb}} \tag{8.279}$$

Using this identity,  $HD_3$  reduces to

$$HD_3 ext{ (level 1)} = \frac{V_{id}^2}{8} \left( K'_{2g_m} \right)^2$$
 (8.280)

Using  $K_{2g_m}'=\frac{1}{2}\left(V_{GS}-V_T\right)$ , we find

$$HD_3 ext{ (level 1)} = \frac{V_{id}^2}{32(V_{GS} - V_T)^2}$$
 (8.281)

which is the same value found in equation (8.255) using the DC transfer characteristic. It is seen that with the level 1 model the bulk effect does not play any role for  $HD_3$  except of course that the bulk effect increases the threshold voltage, such that  $(V_{GS} - V_T)$  is smaller.

The modeling of the bulk effect with the level 1 model is poor, since it does not take into account the variation of the depletion layer width along the channel. Therefore, we check the influence of the bulk effect with a more advanced model. If no bulk effect is present, then  $HD_3$  is found from equation (8.278) by setting  $g_{mb}$  and  $K_{2g_{mb}}^2$  equal to zero:

$$HD_3$$
 (no bulk effect)  $=\frac{1}{8g_m}\cdot\left(\frac{K_{3g_m}}{2}-\frac{K_{2g_m}^2}{g_m}\right)\cdot V_{id}^2$  (8.282)

We now evaluate  $HD_3$  numerically with and without bulk effect using the model of equation (7.121), that takes into account mobility reduction and velocity saturation and the variation of the depletion layer width along the channel. To this purpose, we first bias the transistors  $M_{1A}$  and  $M_{1B}$  with a bulk effect and compute the small-signal parameters and nonlinearity coefficients. Then the transistors are biased without a bulk effect in such a way that the value of  $(V_{GS} - V_T)$  is the same as with the bulk effect. For both cases  $HD_3$  is then evaluated.

The bias conditions and the computed small-signal parameters and nonlinearity coefficients for these two cases are shown in Table 8.4. The model parameters used in the computations are the ones from the  $0.7\mu m$  process of Table 7.1 except that  $v_{sat}=10^5 m/s$  and  $\theta=0.05 V^{-1}$ . These parameters are different since the parameters  $v_{sat}$  and  $\theta$  of Table 7.1 have been fitted for a different model of mobility reduction and velocity saturation.

Without bulk effect, the evaluation of  $HD_3$  with equation (8.282) yields

$$HD_3$$
 (no bulk effect) =  $0.301V_{id}^2$  (8.283)

The largest contribution to this value comes from the term with  $K_{2g_m}$  in equation (8.282). This term is about 22 times larger than the term with  $K_{2g_m}$ . From equation (8.283) we find a third-order intercept point for harmonic distortion of 1.82V.

If we would use the numerical values from Table 8.4 for a common-source amplifier, then, using equation (8.163) we would find an intercept point  $IP_{3h}$  of 4.15V. This is more than two times larger than for a differential pair. This contrasts with the bipolar case where we found that the third-order intercept point for a differential pair is larger than for a common-emitter amplifier. The relatively small value of  $IP_{3h}$  for a MOS differential pair is due to the fact that it is mainly determined by  $K'_{2g_m}$  whereas  $IP_{3h}$  for a common-source amplifier is only determined by  $K'_{3g_m}$  which is very small (with the level 1 model it is even equal to zero).

When the bulk effect is taken into account, the evaluation of  $HD_3$  with equation (8.278) yields

$$HD_3$$
 (with bulk effect) =  $0.296V_{id}^2$  (8.284)

This yields a third-order intercept point for harmonic distortion of 1.84V.

Hence we can conclude that the bulk effect does not influence  $HD_3$  significantly, except of course that it lowers  $(V_{GS} - V_T)$ . When this decrease is compensated, then almost no difference is found with the value of  $HD_3$  in the absence of bulk effect.

parameter	with	without
	bulk effect	bulk effect
W	$160\mu m$	$160\mu m$
L	$0.7\mu m$	$0.7\mu m$
$V_{GS}$	1.25V	1.065V
$V_{SB}$	0.5V	0V
$V_{DS}$	1.45V	1.45V
$V_T$	0.934V	0.75V
$V_{GS} - V_T$	0.316V	0.316V
$i_{DSAT}$	0.730mA	0.70mA
$g_m$	4.57mA/V	4.38mA/V
$K_{2g_m}$	$6.88mA/V^2$	$6.67mA/V^2$
$K_{3g_m}$	$-1.06mA/V^3$	$-0.898mA/V^3$
$g_{mb}$	1.46mA/V	1.75mA/V
$K_{2g_{mb}}$	$-0.927 mA/V^2$	$-1.46mA/V^2$
$K_{3g_{mb}}$	$0.249mA/V^3$	$0.607mA/V^3$
$K_{2_{g_m}\&g_{mb}}$	$-4.28mA/V^2$	$-5.09mA/V^2$
$K_{3_{2g_m\&g_{mb}}}$	$1.32mA/V^3$	$1.69mA/V^3$
$K_{3_{g_m}\&2g_{mb}}$	$0.192mA/V^3$	$0.361mA/V^3$

Table 8.4: Dimensions, bias conditions and the corresponding small-signal parameters and non-linearity coefficients for a transistor with and without bulk effect. The value of  $(V_{GS} - V_T)$  is the same in both cases.

# 8.5.3 $HD_2$ due to mismatches

When there are mismatches between the transistors  $M_{1A}$  and  $M_{1B}$ , then the differential pair no longer operates in a balanced way. As a result, the even-order responses at the output will not be suppressed completely anymore.

According to the work of [Pel 89, Bas 95], the mismatch of two MOS transistors that have

been designed identically, is characterized by the random variation of the difference in their threshold voltage  $V_{TO}$ , their body-effect coefficient  $\gamma$  and their "current factor" that is given by

$$\beta = \mu C_{ox}' \frac{W}{L} \tag{8.285}$$

For illustration purposes, we will only assume a mismatch of 10mV on the threshold voltages. In addition, we will assume that there is a mismatch of 1% between the resistors  $R_{LA}$  and  $R_{LB}$ . With these mismatches, we will now compute the second harmonic distortion due to mismatches in a similar way as we did for the bipolar differential pair in Section 8.4.3. Hereby, we will not take into account bulk effect for simplicity.

To this purpose, we will assume that a circuit element is written as the sum of its nominal value and a mismatch term, as we did in Section 8.4.3. In this way, we have

$$G_{LA} = G_L + \frac{\Delta G_L}{2}$$
  $G_{LB} = G_L - \frac{\Delta G_L}{2}$  (8.286)

$$K_{2g_{m1A}} = K_{2g_m} + \frac{\Delta K_{2g_m}}{2} \qquad K_{2g_{m1B}} = K_{2g_m} - \frac{\Delta K_{2g_m}}{2}$$
 (8.287)

$$g_{m1A} = g_m + \frac{\Delta g_m}{2}$$
  $g_{m1B} = g_m - \frac{\Delta g_m}{2}$  (8.288)

We now evaluate the nominal values and the mismatch terms for two transistors that ideally match in the  $0.7\mu m$  process of Table 7.1. Their parameters are:  $W=80\mu m$ ,  $L=0.7\mu m$ ,  $V_{GS}=1.25V$  and  $V_{SB}=0V$ . The threshold voltage of transistor  $M_{1B}$  is 750mV, which is the nominal value. In this way we find  $V_{GS}-V_T=0.5V$ . The threshold voltage of transistor  $M_{1A}$  has been taken equal to 740mV. The drain current and its derivatives have been computed based upon the model equation (7.121). In this way we find for the nominal drain current

$$i_D = 0.882mA (8.289)$$

For the nominal value of the transconductance  $g_m$  and the second-order nonlinearity coefficient  $K_{2g_m}$  we find

$$g_m = 3.40 \, mA/V$$
  $K_{2g_m} = 3.03 \, mA/V^2$  (8.290)

The mismatch terms are found to be

$$\Delta g_m = 60.7 \mu A/V$$
  $\Delta K_{2g_m} = -18.3 \mu A/V^2$  (8.291)

or, normalized to the nominal values

$$\frac{\Delta g_m}{g_m} = 0.0178 \qquad \frac{\Delta K_{2g_m}}{K_{2g_m}} = -0.00605 \tag{8.292}$$

Clearly, the sign of the mismatch term  $\Delta g_m$  is different from the sign of  $\Delta K_{2g_m}$ . This can be explained by considering the approximate expressions of  $g_m$  and  $K_{2g_m}$  of equations (7.137)

and (7.138), respectively:  $g_m$  is roughly proportional to  $(V_{GS} - V_T)$  whereas  $K_{2g_m}$  is inversely proportional to  $(V_{GS} - V_T)$ .

With the above numerical data we compute the second harmonic distortion. The load resistor  $R_L$  has a value of  $2k\Omega$ . According to equation (8.259) this corresponds to a gain of about 16.6 dB.

The exact expression for the second harmonic distortion has been computed using ISAAC. It consists of 16 terms. An approximate expression consisting of three terms is given by

$$HD_2 \approx \frac{V_{in}}{8} K'_{2g_m} \left( \frac{3\Delta g_m}{g_m} - \frac{\Delta K_{2g_m}}{K_{2g_m}} + \frac{15}{2} \frac{\Delta g_m}{g_m} \frac{\Delta G_L}{G_L} \right)$$
 (8.293)

Only two of these terms contain only one single mismatch term. The third term and the other 13 terms terms that have been neglected contain two or more mismatch terms. The term that contains  $\Delta G_L$  also contains  $\Delta g_m$ . This means that, if no mismatches between the two transistors are present, then a mismatch between the two resistors does not yield a second harmonic. This no longer holds when the output conductance of the two transistors is taken into account.

Further it is seen that  $HD_2$  is determined by mismatches on  $g_m$  and on  $K_{2g_m}$ . In our example these mismatches are correlated, since they both originate from a mismatch in the threshold voltage. Since  $\Delta g_m$  and  $\Delta K_{2g_m}$  have an opposite sign, the terms with  $\Delta g_m$  and  $\Delta K_{2g_m}$  add. On the other hand, the sign of the third term is not fixed with respect to the first two terms, since the third term is due to mismatches on the load resistance as well, which are uncorrelated with respect to the  $V_T$  mismatches.

It is interesting to compare the absolute values of the three terms of  $HD_2$  in equation (8.293). With our numerical example we find that the term  $3\Delta g_m/g_m$  is about nine times larger than the term  $\Delta K_{2g_m}/K_{2g_m}$ . The third term is in turn 4.5 times smaller than the term in  $\Delta K_{2g_m}/K_{2g_m}$ . Finally, if the sign of the third term is taken such that it has the same sign as the first two terms then we find for  $HD_2$ 

$$HD_2 = 6.77 \times 10^{-3} V_{in} \tag{8.294}$$

This yields an intercept point  $IP_{2h}$  of 147V. This is nearly two orders of magnitude larger than for the third-order intercept point  $IP_{3h}$  which is not determined by mismatches.

Finally, it can be remarked that the first two terms in equation (8.293) are identical to the dominant terms of  $HD_2$  of a bipolar differential pair.

## 8.6 Emitter follower

Figure 8.47 depicts an emitter follower. The output of this circuit is at the emitter of transistor  $Q_1$ . Since an emitter follower is a circuit with local negative feedback, it is expected that the distortion will be suppressed by the loop gain of this feedback.

We will analyze the second and third harmonic distortion of the emitter follower starting from the nonlinear circuit of Figure 8.48. This circuit contains two nonlinearities: the base current and the collector current. Both currents are a function of the base-emitter voltage.

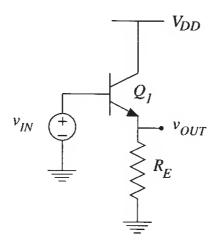


Figure 8.47: An emitter follower.

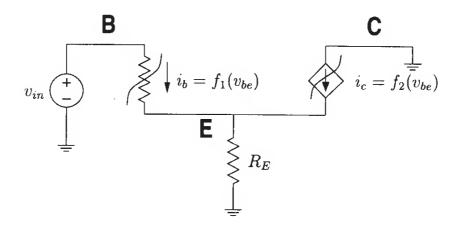


Figure 8.48: AC-equivalent circuit of the emitter follower.

## 8.6.1 First-order response

The first-order response is found by linearizing the circuit of Figure 8.48. This yields the circuit of Figure 8.49.

The output of interest in this circuit is the response  $V_{e,1,0}$  at the emitter of the transistor. Using simple network analysis we find

$$V_{e,1,0} = \frac{g_m + g_\pi}{g_m + g_\pi + G_E} \cdot V_{in}$$
 (8.295)

It is seen that, when  $G_E = 0$ , then  $V_{e,1,0} = V_{in}$ . This situation corresponds to the case where  $R_E$  corresponds to an ideal current source. In practice,  $G_E$  is always larger than zero due to the presence of the output conductance of the transistor which has been neglected in this analysis.

For further computations we will need the first-order response of the base-emitter voltage. The reason is that the nonlinear base current and the nonlinear collector current are both controlled by the base-emitter voltage. Hence, the first-order response  $V_{be,1,0}$  will determine the

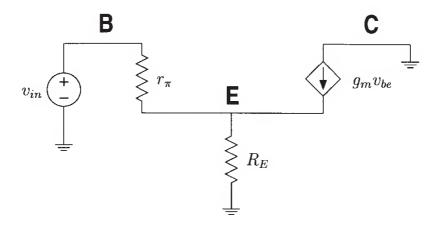


Figure 8.49: Linearized equivalent of the circuit of Figure 8.48.

nonlinear current sources that correspond to the nonlinearity of the collector and the base current.

For  $V_{be,1,0}$  we find, using equation (8.295)

$$V_{be,1,0} = V_{in} - \frac{g_m + g_\pi}{g_m + g_\pi + G_E} \cdot V_{in} = \frac{G_E}{g_m + g_\pi + G_E} \cdot V_{in}$$
(8.296)

When  $G_E$  is zero, then it is seen that  $V_{be,1,0}$  is zero as well. This corresponds to an infinite log gain of the internal feedback of the emitter follower.

## 8.6.2 Second-order response

For the computation of the second-order response, the linearized circuit is excited by two nor linear current sources of order two, one corresponding to the base current nonlinearity and or corresponding to the collector current nonlinearity. This situation is depicted in Figure 8.50.

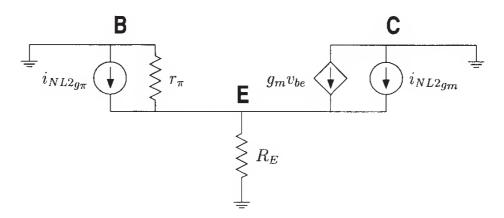


Figure 8.50: Linearized equivalent of the circuit of Figure 8.48 excited with the nonlinear current sources of order two.

The two current sources appear to be in parallel. Hence they can be combined to one single source  $i_{NL2tot}$ . Using Table 5.5 we find

$$i_{NL2tot} = \frac{K_{2g_m} + K_{2g_{\pi}}}{2} \cdot V_{be,1,0}^2$$
 (8.297)

Using equation (8.296) we obtain

$$i_{NL2tot} = \frac{1}{2} \left( K_{2g_m} + K_{2g_{\pi}} \right) \frac{G_E^2}{\left( g_m + g_{\pi} + G_E \right)^2} \cdot V_{in}^2$$
 (8.298)

The second harmonic at the output, which is the node voltage at the emitter, is found using simple network analysis in the circuit of Figure 8.50. One finds

$$V_{e,2,0} = \frac{i_{NL2tot}}{g_m + g_\pi + G_E} \tag{8.299}$$

Substituting the value of  $i_{NL2tot}$  from equation (8.298),  $V_{e,2,0}$  becomes

$$V_{e,2,0} = \frac{1}{2} \frac{\left(K_{2g_m} + K_{2g_{\pi}}\right) G_E^2}{\left(g_m + g_{\pi} + G_E\right)^3} \cdot V_{in}^2$$
(8.300)

The second harmonic distortion is found by dividing  $V_{e,2,0}$  by  $V_{e,1,0}$ . Doing so, we obtain after some rearrangements

$$HD_2 = \frac{1}{2} \frac{\left(K_{2g_m} + K_{2g_\pi}\right)}{\left(g_m + g_\pi\right)\left(1 + \left(g_m + g_\pi\right)R_E\right)^2} \cdot V_{in}$$
(8.301)

For a realistic transistor beta, say 40 or more,  $K_{2g_{\pi}}$  is negligible compared to  $K_{2g_m}$ , and the same holds for  $g_{\pi}$  versus  $g_m$ . Then  $HD_2$  reduces to

$$HD_2 \approx \frac{1}{2} \frac{K'_{2g_m}}{(1 + g_m R_E)^2} \cdot V_{in}$$
 (8.302)

The factor  $g_m R_E$  is the loop gain of the local feedback of the emitter follower [Gray 93]. Hence we see that the second harmonic distortion is divided by the square of the loop gain, as we found in Section 4.8.

The second harmonic distortion is seen to be a factor  $1/(g_m R_E)^2$  smaller than the second harmonic distortion of a one transistor amplifier. For this circuit we found a second-order intercept point  $IP_{2h}$  of 103mV. Assume now that the collector current through  $Q_1$  is 1mA and  $R_E = 1k\Omega$ . Then  $g_m R_E$  is 38.76, and so we find  $IP_{2h} \approx 155V$ .

For further computations we will need the second-order response of the base-emitter voltage. The base is fixed to the external voltage source. Since this is assumed to be a pure sinusoidal signal without harmonics, there are no harmonics at the base of  $Q_1$ . Hence

$$V_{be,2,0} = V_{b,2,0} - V_{e,2,0} = 0 - V_{e,2,0} = -V_{e,2,0}$$
(8.303)

and  $V_{e,2,0}$  is given in equation (8.300).

#### 8.6.3 Third-order response

For the computation of the third harmonic, the linearized circuit of Figure 8.49 is excited with nonlinear current sources of order three. The situation is identical to what is shown in Figure 8.50, except that the index "2" of the nonlinear current sources must be replaced by "3". From Table 5.7 we find that the sum of the nonlinear current sources that correspond to the base current nonlinearity and to the collector current nonlinearity, is given by

$$i_{NL3tot} = \left(K_{2g_m} + K_{2g_{\pi}}\right) V_{be,1,0} V_{be,2,0} + \frac{1}{4} \left(K_{3g_m} + K_{3g_{\pi}}\right) V_{be,1,0}^3$$
(8.304)

The third-order response at the emitter is found in the same way as for order two:

$$V_{e,3,0} = \frac{i_{NL3tot}}{g_m + g_\pi + G_E} \tag{8.305}$$

Using equations (8.296), (8.300), (8.303) and (8.304) we find

$$V_{e,3,0} = \frac{G_E^3}{4\left(g_m + g_\pi + G_E\right)^5} \cdot \left[ \left( K_{3g_m} + K_{3g_\pi} \right) \left( g_m + g_\pi + G_E \right) - 2\left( K_{2g_m} + K_{2g_\pi} \right)^2 \right] \cdot V_{in}^3$$
(8.306)

The third harmonic distortion is found by dividing  $V_{e,3,0}$  by  $V_{e,1,0}$ . This yields

$$HD_{3} = \frac{G_{E}^{3} \cdot V_{in}^{2}}{4(g_{m} + g_{\pi}) (g_{m} + g_{\pi} + G_{E})^{4}} \cdot \left[ \left( K_{3g_{m}} + K_{3g_{\pi}} \right) (g_{m} + g_{\pi} + G_{E}) - 2 \left( K_{2g_{m}} + K_{2g_{\pi}} \right) \right]$$
(8.30)

Clearly,  $HD_3$  consists of two terms with an opposite sign. One can compute that the two terms compensate each other for  $g_m R_E \approx 0.5$ . Compared to the results of Section 4.8.3.2, obtains for a one-transistor amplifier with emitter degeneration, it is seen that  $HD_3$  becomes zero for the same value of the loop gain.

Assume now that  $\beta$  is large and that  $g_m R_E \gg 1$ . Then the two terms inside the squabrackets of equation (8.307) reduce to  $(K_{3g_m}g_m - 2K_{2g_m}^2)$ . Using the simple expressions  $K_{2g_m}$  and  $K_{3g_m}$  given in equations (3.14) and (3.15), we find for  $HD_3$ 

$$HD_3$$
(large loop gain)  $\approx \frac{1}{12} \cdot \frac{1}{(g_m R_E)^3} \cdot \frac{V_{in}^2}{V_t^2}$  (8.30)

The third-order intercept for harmonic distortion is found from equation (8.308) by setting  $HD_3$  equal to one and solving for  $V_{in}$ . This yields

$$IP_{3h} = 2\sqrt{3} V_t (g_m R_E)^{3/2} (8.36)$$

For a collector current of 1mA and  $R_E=1k\Omega$  we find  $g_mR_E=38.76$  such that  $IP_{3h}=21.6$ 

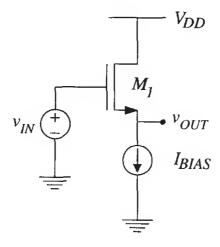


Figure 8.51: A source follower.

#### 8.7 Source follower

Figure 8.51 shows a source follower. Compared to the bipolar case in the previous section, the MOS transistor here is loaded with an ideal current source. Nevertheless, the gain of the source follower is not one, as we would expect from the bipolar case. Indeed, we will see that due to the bulk effect the small-signal gain is lower than one. Since this bulk effect is nonlinear, it will influence the harmonic distortion at the source node.

For the computation of the harmonic distortion with the method of Section 5.3, we start from the circuit of Figure 8.52, that is AC equivalent to the source follower of Figure 8.51.

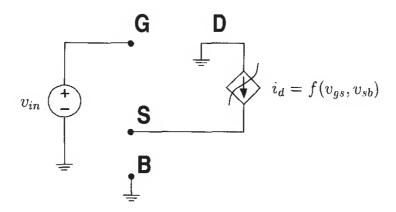


Figure 8.52: AC-equivalent circuit of the source follower of Figure 8.51.

In the circuit of Figure 8.52 the drain current is represented as a nonlinear function of the gate-source and the source-bulk voltage. The dependence on the drain-source voltage has been omitted. This means that we will neglect the output conductance of the transistor.

#### 8.7.1 First-order response

The linearized equivalent of the nonlinear circuit of Figure 8.52 is shown in Figure 8.53. It is

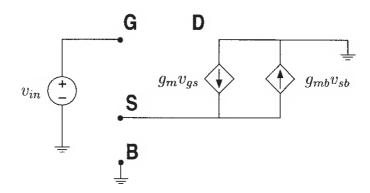


Figure 8.53: Linearized equivalent circuit of Figure 8.52.

seen that the current generator  $g_{mb} \cdot v_{sb}$  is controlled by the voltage over its own terminals. Hence it can be regarded as a resistor with a value  $1/g_{mb}$ . From the linearized circuit we easily compute the first-order response at the source node:

$$V_{s,1,0} = \frac{g_m}{g_m + g_{mb}} \cdot V_{in} \tag{8.310}$$

If we compare this response to the output of the emitter follower from Figure 8.47 where  $g_{\pi}$  of transistor  $Q_1$  is made equal to zero, then it is seen that  $g_{mb}$  plays the same role as the conductance  $G_E$  at the emitter of  $Q_1$ . Since  $g_{mb} > 0$  we find from equation (8.310) that  $V_{s,1,0} < V_{in}$ .

The first-order response of the gate-source voltage is given by

$$V_{gs,1,0} = V_{in} - V_{s,1,0} = \frac{g_{mb}}{g_m + g_{mb}} \cdot V_{in}$$
(8.311)

This response will be used for the calculation of the higher-order responses.

# 8.7.2 Second-order response

For the computation of the second harmonic the linearized circuit is excited with nonlinear current sources of order two. Since in Figure 8.52 the drain current has been modeled as a two dimensional nonlinearity, there are several nonlinear second-order current sources in parallel between the drain and the source. We have combined them into one single source  $i_{NL2tot}$  with three components. These correspond to the coefficients  $K_{2g_m}$ ,  $K_{2g_{mb}}$  and  $K_{2g_m\&g_{mb}}$ , respectively.

The component that is proportional to  $K_{2g_m}$  is given by (see Table 5.5)

$$i_{NL2gm} = \frac{K_{2g_m}}{2} \cdot V_{gs,1,0}^2 = \frac{K_{2g_m} g_{mb}^2}{2(g_m + g_{mb})^2} \cdot V_{in}^2$$
(8.312)

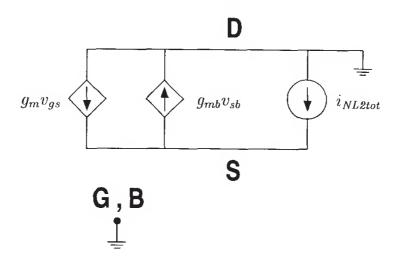


Figure 8.54: Linearized equivalent circuit of Figure 8.52 excited with a nonlinear current source of order two.

The component that is proportional to  $K_{2g_{mb}}$  does not flow from the drain to the source but from the source to the drain. This is due to the convention we made about the derivatives of the drain current and its derivatives with respect to  $v_{SB}$ . In Sections 3.2.5 and 7.2 it was stated that  $g_{mb}$  and the nonlinearity coefficients  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  are proportional to the derivatives of the drain current with respect to  $v_{SB}$ , multiplied with -1. This yields a positive value for  $g_{mb}$ . As a result, the current  $g_{mb}v_{sb}$  flows from source to drain. Similarly, the orientation of the nonlinear current sources of order two and three that correspond to  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$ , is from source to drain. Hence, using Table 5.5, the component of  $i_{NL2tot}$  that is proportional to  $K_{2g_{mb}}$  is given by

$$-i_{NL2g_{mb}} = -\frac{K_{2g_{mb}}}{2} V_{s,1,0}^2 = -\frac{K_{2g_{mb}} g_m^2}{2(g_m + g_{mb})^2} \cdot V_{in}^2$$
(8.313)

Finally, the component that is proportional to  $K_{2g_m\&g_{mb}}$  is found from Table 5.5 to be

$$i_{NL2_{gm}\&g_{mb}} = \frac{K_{2_{g_m\&g_{mb}}}}{2} \cdot V_{gs,1,0} \cdot V_{s,1,0} = \frac{K_{2_{g_m\&g_{mb}}}g_mg_{mb}}{2\left(g_m + g_{mb}\right)^2} \cdot V_{in}^2$$
(8.314)

The second-order response is found using simple network analysis in Figure 8.54:

$$V_{s,2,0} = \frac{i_{NL2tot}}{g_m + g_{mb}} \tag{8.315}$$

Substituting the different components of  $i_{NL2tot}$  into equation (8.315) yields

$$V_{s,2,0} = \frac{V_{in}^2}{2(g_m + g_{mb})^3} \cdot \left( K_{2g_m} g_{mb}^2 - K_{2g_{mb}} g_m^2 + K_{2g_m \& g_{mb}} g_m g_{mb} \right)$$
(8.316)

The second harmonic distortion is found as the ratio of  $V_{s,2,0}$  and  $V_{s,1,0}$ :

$$HD_2 = \frac{V_{in}}{2(g_m + g_{mb})^2 g_m} \cdot \left| K_{2g_m} g_{mb}^2 - K_{2g_{mb}} g_m^2 + K_{2g_m \& g_{mb}} g_m g_{mb} \right|$$
(8.317)

The second-order intercept point  $IP_{2h}$  for harmonic distortion can be found by setting  $HD_2$  equal to one and solving for  $V_{in}$ . Doing so for the transistor with the parameters of the left column in Table 8.4 we find  $IP_{2h} = 61V$ . This is a smaller value than we found in the example of the emitter follower in Section 8.6.2. This can be explained partially by the smaller value of the loop gain of the internal feedback, which is  $(g_m/g_{mb})$ , compared to the value  $g_mR_E$  for an emitter follower.

It is interesting to compare the values of the different terms of  $HD_2$  in equation (8.317). With the numerical values of the left column of Table 8.4 the value of the term with  $K_{2g_m}$  is  $0.0442 \cdot V_{in}$ , the term with  $K_{2g_{mb}}$  equals  $0.0583 \cdot V_{in}$  and the term with  $K_{2g_{mb}}$  is  $-0.0861 \cdot V_{in}$ . Hence, the term with  $K_{2g_m}$  is the smallest term. This means that in this example the nonlinearity of the bulk effect plays a larger role than the nonlinearity of the drain current dependence on  $v_{GS}$ . Further it is seen that the three nonlinearities partially compensate since they do not all have the same sign.

Assume now that  $K_{2g_{mb}}$  and  $K_{2g_{mb}g_{mb}}$  are zero. In this case,  $HD_2$  is given by the first term in the right-hand side of equation (8.317):

$$HD_2 = \frac{K_{2g_m} g_{mb}^2}{2(g_m + g_{mb})^2 g_m} \cdot V_{in}$$
(8.318)

Now we have a similar expression as for  $HD_2$  at the output of an emitter follower (see equation (8.302)). Indeed, the transconductance  $g_{mb}$  in the source follower plays the same role as the conductance  $G_E=1/R_E$  in the emitter follower of Figure 8.47. Finally, it must be remarked that  $HD_2$  given in equation (8.318) is larger than  $HD_2$  from equation (8.317), where the nonlinearity of the bulk effect has been taken into account. Indeed, with equation (8.318) we find  $IP_{2h}=22.6V$ . This is due to the fact that there is no compensating effect anymore from the nonlinearity coefficients  $K_{2g_{mb}}$  and  $K_{2g_{mb}\& g_{mb}}$ .

Inclusion of the output conductance The expression for  $HD_2$  of equation (8.317) can be extended to include the output conductance and its nonlinearity. Indeed, the linearized output conductance  $g_o$  appears to be in parallel with  $g_{mb}$  in the source follower circuit. Hence  $g_o$  can simply be added to  $g_{mb}$ .

Further, the nonlinear current sources of order two that arise from the output conductant nonlinearity all flow from the drain to the source. The nonlinear current source that is proportional to  $K_{2g_o}$  is determined by  $V_{ds,1,0}^2$ . Since in the source follower circuit  $V_d = V_b$ , we fit that  $V_{ds,1,0}^2 = V_{sb,1,0}^2$ . Hence, the contribution of  $K_{2g_o}$  can simply be added to the contribution of  $K_{2g_{mb}}$ . However, the two contributions have an opposite sign due to the convention we man about the orientation of the nonlinear current source that corresponds to  $K_{2g_{mb}}$ . For the nonlinear current sources that correspond to the coefficients  $K_{2g_m\&g_o}$  and  $K_{2g_{mb}\&g_o}$  a similar reasoning can be made. Finally we obtain for the second harmonic distortion including the output conductance nonlinearity

$$HD_{2} = \frac{V_{in}}{2(g_{m} + g_{mb} + g_{o})^{2}g_{m}} \cdot \left(K_{2g_{m}} (g_{mb} + g_{o})^{2} + \left(K_{2g_{o}} - K_{2g_{mb}} - K_{2g_{mb}\&g_{o}}\right)g_{m}^{2} + \left(K_{2g_{m}\&g_{mb}} - K_{2g_{m}\&g_{o}}\right)g_{m} (g_{mb} + g_{o})\right)$$
(8.31)

8.7 Source follower

Nonlinear emitter degeneration Assume now that  $K_{2g_{mb}}$  differs from zero while  $K_{2g_{m}\&g_{mb}}$  is zero. In this case  $K_{2g_m}$  and  $K_{2g_{mb}}$  can correspond to two independent second-order nonlinearities that do not arise from a Taylor series expansion of a multi-dimensional nonlinearity. In this way, the coefficient  $K_{2g_{mb}}$  can be identified with a nonlinear conductance between the source node of the transistor and ground, while  $K_{2g_m}$  models the nonlinear dependence of the transistor current on its controlling voltage  $v_{GS}$ . This situation can be identified with the situation of a MOS source follower with the bulk connected to the source. Alternatively, this situation also corresponds to a bipolar emitter follower in which the base current is neglected and which is loaded at its emitter with a nonlinear conductance. If in equation (8.317) we replace  $g_{mb}$  by a conductance  $G_E$  and  $K_{2g_{mb}}$  by  $K_{2G_E}$ , which is the second-order nonlinearity coefficient of the nonlinear conductance between the emitter and ground, then we find

$$HD_2 = \frac{V_{in}}{2(g_m + G_E)^2 g_m} \cdot \left( K_{2g_m} G_E^2 - K_{2G_E} g_m^2 \right)$$
 (8.320)

From this equation it is seen that  $HD_2$  can be made zero if

$$\frac{K_{2g_m}}{K_{2G_E}} = \sqrt{\frac{g_m}{G_E}} \tag{8.321}$$

If this condition is met, then in reality  $HD_2$  will not be zero due to other effects that have been neglected here, such as the base current.

In the case of a MOS source follower that is degenerated with a nonlinear load conductance described by coefficients  $G_E$  and  $K_{2G_E}$ , the conductance  $G_E$  simply adds to  $g_{mb}$  and  $K_{2G_E}$  to  $K_{2g_{mb}}$ .

# 8.7.3 Third-order response

For the computation of the third harmonic the linearized circuit of Figure 8.53 is excited with nonlinear current sources of order three. The situation is identical to what is shown in Figure 8.54 except that the second-order nonlinear current source  $i_{NL2tot}$  must be replaced by a third-order one, which we denote by  $i_{NL3tot}$ . This current source consists of the components caused by the nonlinearity coefficients  $K_{2g_m}$ ,  $K_{3g_m}$ ,  $K_{2g_{mb}}$ ,  $K_{3g_{mb}}$ ,  $K_{2g_m\&g_{mb}}$  and  $K_{3g_m\&2g_{mb}}$ . Similarly to what we found in the previous section for the second-order response, the third-

Similarly to what we found in the previous section for the second-order response, the third-order response is given by

$$V_{s,3,0} = \frac{i_{NL3tot}}{g_m + g_{mb}} \tag{8.322}$$

and  $HD_3$  is given by

$$HD_3 = \frac{V_{s,3,0}}{V_{s,1,0}} = \frac{i_{NL3tot}}{g_m V_{in}}$$
 (8.323)

The value of  $i_{NL3tot}$  can be found using Table 5.7. The resulting expression is quite lengthy. The different terms of the numerator of  $HD_3$  are shown in Figure 8.55 together with their numerical

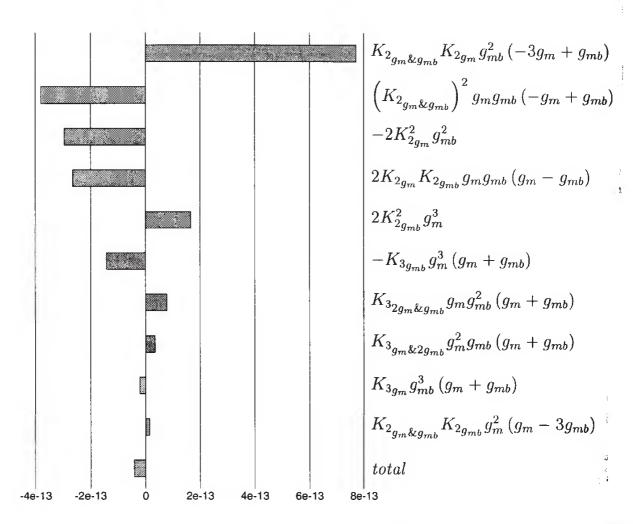


Figure 8.55: The different contributions to the numerator of  $HD_3$  at the output of the source follower of Figure 8.51.

value obtained for the parameters of the middle column in Table 8.4. The denominator of *HB* is given by

$$4g_m \left(g_m + g_{mb}\right)^4 \tag{8.324}$$

It is seen from Figure 8.55 that the total value is much smaller in absolute value than the largest term due to the compensating effect of the different terms. An easy interpretation is not possible here. A numerical evaluation of the different terms yields a third-order intercept point  $IP_{3h}$  of 24.3V.

In order to include the output conductance and its nonlinearity into the expression of  $HD_3$  the same approach can be followed as for order two.

Next, we can again consider some special cases as we did for order two. Assume first that the bulk effect is linear, such that  $K_{2g_{mb}} = K_{2g_m\&g_{mb}} = 0$  and  $K_{3g_{mb}} = K_{3g_m\&g_{mb}} = K_{3g_m\&2g_{mb}} = 0$ . In this case, the numerator of  $HD_3$  given in Figure 8.55 together with the denominator

equation (8.324) reduce to

$$HD_{3} = \frac{-2K_{2g_{m}}^{2}g_{mb}^{3} + K_{3g_{m}}g_{mb}^{3}(g_{m} + g_{mb})}{4(g_{m} + g_{mb})^{4}g_{m}} \cdot V_{in}^{2}$$
(8.325)

Now we have again the same expression as for the third harmonic distortion of an emitter follower (see equation (8.307)) except that the base current is not considered here. Indeed,  $g_{mb}$  plays the same role as the emitter degeneration conductance  $G_E = 1/R_E$  of Figure 8.47.

Next, we consider the case where all nonlinearity coefficients that correspond to cross-derivatives of the drain current with respect to  $v_{GS}$  and  $v_{SB}$  are all zero. In this case the bulk transconductance reduces to a nonlinear conductance between the source node of the transistor and ground, while  $K_{2g_m}$  and  $K_{3g_m}$  model the nonlinear dependence of the transistor current on its controlling voltage  $v_{GS}$ . This situation can be identified with the situation of a MOS source follower with the bulk connected to the source. Alternatively, this situation also corresponds to a bipolar emitter follower in which the base current is neglected and which is loaded at its emitter with a nonlinear conductance. If we replace  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  by the coefficients  $G_E$ ,  $K_{2G_E}$  and  $K_{3G_E}$ , that describe a nonlinear conductance, then we find from the terms in Figure 8.55

$$HD_{3} = \left[ -K_{3_{G_{E}}} g_{m}^{3} \left( g_{m} + G_{E} \right) + 2K_{2g_{m}} K_{2_{G_{E}}} g_{m} G_{E} \left( g_{m} - G_{E} \right) \right]$$
(8.326)

$$+2K_{2G_{E}}^{2}g_{m}^{3}-2K_{2g_{m}}^{2}G_{E}^{3}+K_{3g_{m}}G_{E}^{3}\left(g_{m}+G_{E}\right)$$
  $\left[4\left(g_{m}+G_{E}\right)^{4}g_{m}\right]$  (8.327)

Again, we can derive conditions for the nonlinearity coefficients of the load conductance such that  $HD_3$  is zero. These conditions could be combined with the condition given in equation (8.321), such that both  $HD_2$  and  $HD_3$  are zero for an emitter follower with a nonlinear load.

# 8.8 Cascode transistor

In this section we check whether a cascode transistor causes distortion when it is driven by a current. Ideally, a cascode transistor simply passes the input current to the load at its drain (MOS transistor) or collector (bipolar transistor) without any loss and without distortion. In fact, a cascode transistor acts as a common-base (bipolar) or a common-gate (MOS) stage. The study of such stage that is driven by a voltage is postponed to the next section.

#### 8.8.1 Bipolar cascode

With a bipolar cascode transistor some nonlinear distortion will arise even when it is driven by an ideal current source, as shown in Figure 8.56. The reason is that a part of the current flows to the ground through  $r_{\pi}$ . In addition,  $r_{\pi}$  is nonlinear.

We will now compute an expression for the second and third harmonic distortion of the current through the resistance  $R_L$ . Hereby we will neglect capacitors, ohmic resistors as well as the output conductance of the transistor.

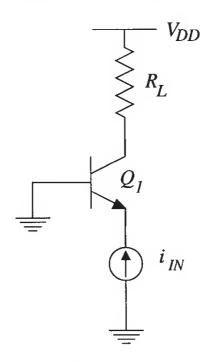


Figure 8.56: A bipolar cascode transistor driven by a current source.

The nonlinear circuit which is equivalent to the cascode stage of Figure 8.56 is shown in Figure 8.57. This circuit takes into account two nonlinearities, namely the base current and the

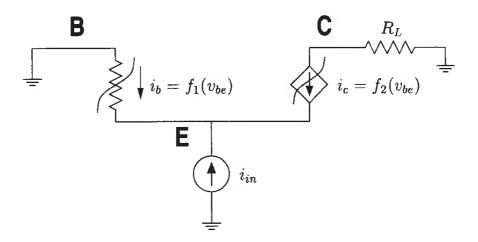


Figure 8.57: AC-equivalent circuit of the cascode stage of Figure 8.56.

collector current, which are both functions of the base-emitter voltage.

#### 8.8.1.1 First-order response

First, we compute the fundamental of the current through  $R_L$ . To this end, the circuit of Figure 8.57 is linearized. This yields the circuit of Figure 8.58.

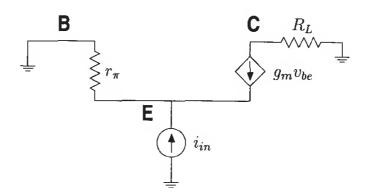


Figure 8.58: Linearized equivalent of the cascode stage of Figure 8.58.

In this circuit we easily find the voltage at the emitter of  $Q_1$ :

$$V_{e,1,0} = \frac{I_{in}}{g_m + g_{\pi}} \tag{8.328}$$

Since the base is grounded, the base-emitter voltage is found to be

$$V_{be,1,0} = 0 - V_{e,1,0} = -\frac{I_{in}}{q_m + q_{\pi}}$$
(8.329)

The current through  $R_L$  is then given by

$$I_{out,1,0} = -g_m V_{be,1,0} = \frac{g_m}{g_m + g_\pi} \cdot I_{in}$$
 (8.330)

#### 8.8.1.2 Second-order response

In order to find the second-order component  $I_{out,2,0}$  of the current through  $R_L$ , the linearized circuit is excited with the nonlinear current sources of order two, as shown in Figure 8.59. One

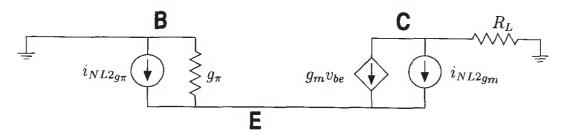


Figure 8.59: Linearized equivalent of the cascode stage of Figure 8.57 excited with nonlinear surrent sources of order two.

ionlinear current source originates from the nonlinearity of the base current, the other one from

the nonlinearity of the collector current. The second-order response at the emitter is found from Figure 8.59:

$$V_{e,2,0} = \frac{i_{NL2g_m} + i_{NL2g_\pi}}{g_m + g_\pi} \tag{8.331}$$

The nonlinear current sources are given by

$$i_{NL2gm} = \frac{K_{2g_m}}{2} \cdot V_{be,1,0}^2 = \frac{K_{2g_m}}{2} \cdot \frac{I_{in}^2}{(q_m + q_\pi)^2}$$
 (8.332)

$$i_{NL2g_{\pi}} = \frac{K_{2g_{\pi}}}{2} \cdot V_{be,1,0}^2 = \frac{K_{2g_{\pi}}}{2} \cdot \frac{I_{in}^2}{(g_m + g_{\pi})^2}$$
(8.333)

Then we find for  $V_{e,2,0}$ 

$$V_{e,2,0} = \frac{K_{2g_m} + K_{2g_{\pi}}}{2} \cdot \frac{I_{in}^2}{(q_m + q_{\pi})^3}$$
(8.334)

and for  $V_{be,2,0}$ 

$$V_{be,2,0} = -V_{e,2,0} = -\frac{K_{2g_m} + K_{2g_{\pi}}}{2} \cdot \frac{I_{in}^2}{(g_m + g_{\pi})^3}$$
(8.335)

From Figure 8.59 we see that the second harmonic  $I_{out,2,0}$  of the output current is the sum of the current of the generator  $g_m \cdot V_{be,2,0}$  and the current of the nonlinear current source of order two that corresponds to the collector current nonlinearity:

$$I_{out,2,0} = i_{NL2gm} + g_m V_{be,2,0} (8.336)$$

Using equations (8.332) and (8.335) we find after some algebra

$$I_{out,2,0} = \frac{I_{in}^2}{2(g_m + g_\pi)^3} \cdot \left( K_{2g_m} g_\pi - K_{2g_\pi} g_m \right) \tag{8.337}$$

The second harmonic distortion is found by dividing  $I_{out,2,0}$  by  $I_{out,1,0}$ . This yields

$$HD_2 = \frac{I_{in}}{2q_m (q_m + q_\pi)^2} \left( K_{2g_m} g_\pi - K_{2g_\pi} g_m \right) \tag{8.338}$$

This second harmonic distortion is seen to consist of two components of opposite sign: one contribution comes from the collector current nonlinearity, the other from the base current nonlinearity. The two contributions cancel if

$$\frac{K_{2g_m}}{K_{2g_\pi}} = \frac{g_m}{g_\pi} \tag{8.339}$$

This occurs when the two nonlinearities track, which is the case if the transistor beta is constant. In reality the beta is current dependent such that the two nonlinearities do not track. With the data from Table 8.2 equation (8.338) yields a second-order intercept point  $IP_{2h}$  of 1.32V.

Assume now that  $r_{\pi}$  (or  $g_{\pi}$ ) is linear. Then we see from equation (8.338) that distortion still occurs. The presence of a linear  $r_{\pi}$  is similar to the presence of a finite linear output resistance of the input current source.

#### 8.8.1.3 Third-order response

For the computation of the third-order response, the linearized circuit of Figure 8.58 is excited with nonlinear current sources of order three. The calculation of this response is left to the reader. The third-order harmonic distortion that results from this calculation is given by

$$HD_{3} = \frac{I_{in}^{2}}{4g_{m}(g_{m} + g_{\pi})^{4}} \left[ K_{3g_{m}} g_{\pi}(g_{m} + g_{\pi}) - K_{3g_{\pi}} g_{m}(g_{m} + g_{\pi}) - 2K_{2g_{m}}^{2} g_{\pi} + 2K_{2g_{m}} K_{2g_{\pi}} (g_{m} - g_{\pi}) \right]$$

$$+2K_{2g_{\pi}}^{2} g_{m} + 2K_{2g_{m}} K_{2g_{\pi}} (g_{m} - g_{\pi})$$

$$(8.340)$$

When the nonlinearities of the base current and the collector current track then

$$\frac{g_m}{g_\pi} = \frac{K_{2g_m}}{K_{2g_\pi}} = \frac{K_{3g_m}}{K_{3g_\pi}} = \beta_F \tag{8.341}$$

In this case, equation (8.340) reduces to

$$HD_{3} = \frac{I_{in}^{2}}{4g_{m}^{5}(1+\frac{1}{\beta_{F}})^{4}} \left[ K_{3g_{m}}g_{m}^{2} \left( \frac{\beta_{F}+1}{\beta_{F}^{2}} \right) - K_{3g_{m}}g_{m}^{2} \left( \frac{\beta_{F}+1}{\beta_{F}^{2}} \right) - 2K_{2g_{m}}^{2} \frac{g_{m}}{\beta_{F}} + 2K_{2g_{m}}^{2} \frac{g_{m}}{\beta_{F}^{2}} - 2K_{2g_{m}}^{2} \frac{g_{m}}{\beta_{F}^{2}} \right] = 0$$

$$(8.342)$$

In reality, the nonlinearities do not track completely. As a result, a small third harmonic occurs. With the data from Table 8.2, equation (8.340) yields a third-order intercept point  $IP_{3h}$  of 0.62V.

#### 8.8.2 MOS cascode

Figure 8.60 depicts a MOS cascode transistor that is driven by a current source. The AC-equivalent circuit is shown in Figure 8.61. From this figure it is seen that the current through  $R_L$  must be identical to the input current, since the input current cannot flow to somewhere else. As a result, no distortion will be produced. Of course, this is an idealization. In reality, the current source has an output conductance that is not zero. As a result, nonlinear distortion will arise.

# 8.9 Common-gate and common-base transistor

Figure 8.62 depicts a common-gate MOS transistor that is driven by a voltage source. The gain of this stage is about the same as the gain of a common-source stage [Gray 93, Lak 94]. The difference with a common-source stage is that bulk effect plays a role here. Now we will investigate whether this common-gate stage produces more or less distortion at its output than a common-source stage.

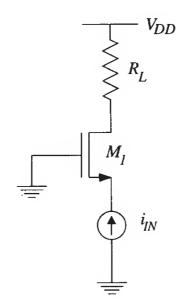


Figure 8.60: A MOS cascode transistor driven by a current source.

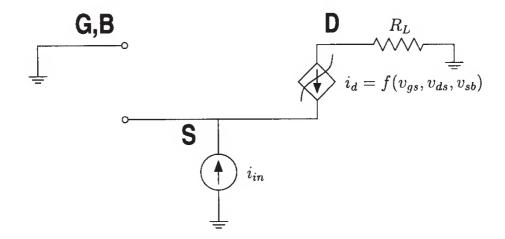


Figure 8.61: AC-equivalent circuit of the cascode stage of Figure 8.60.

In parallel with the analysis of a MOS common-gate stage we will also analyze a bipolar common-base stage, that is depicted in Figure 8.63. The distortion formulas for a common-base stage will be derived from the results obtained for a MOS common-gate stage by omitting all coefficients that model the bulk effect. In the resulting expressions the effect of the base current of the bipolar transistor will of course be neglected.

In order to compute the different harmonics of a common-gate stage we start from the circuit of Figure 8.64 that is the AC-equivalent of the circuit from Figure 8.62. This circuit contains one single nonlinearity, namely the drain current. This current is modeled here as a function of two voltages, namely the gate-source and the source-bulk voltage. The dependence on the drain-source voltage has been omitted in this analysis. The output that is considered here is the current through the load resistance  $R_L$ .

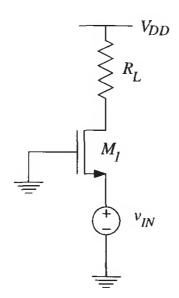


Figure 8.62: A MOS common-gate transistor excited with a voltage source.

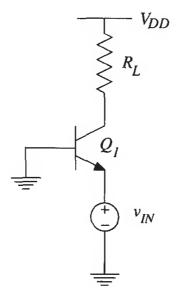


Figure 8.63: A bipolar common-base transistor excited with a voltage source.

# 8.9.1 First-order response

The fundamental component  $I_{out,1,0}$  of the output current is computed with the network of Figure 8.65, which is the linearized equivalent of the circuit of Figure 8.64.

Using simple network analysis, the output current  $I_{out,1,0}$  is found to be

$$I_{out,1,0} = (g_m + g_{mb})V_{in} (8.343)$$

In the computations of the higher-order responses we will need the response at the source node, since the source-bulk and the gate-source voltages control the drain current nonlinearity. It

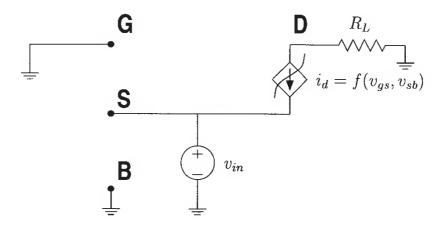


Figure 8.64: AC-equivalent circuit of the common-gate stage of Figure 8.62.

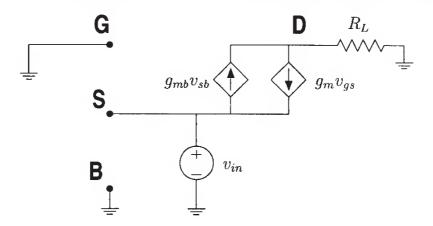


Figure 8.65: Linearized equivalent of the circuit of Figure 8.64.

is seen in Figure 8.65 that the source voltage is simply equal to the input voltage.

#### 8.9.2 Second-order response

For the computation of the second-order response the external excitation is removed. Instead, the nonlinear current sources of order two are applied. This yields the situation of Figure 8.66.

There are three nonlinear current sources of order two:

- a source determined by  $K_{2g_m}$ . This models the dependence of the drain current on  $v_{\mathcal{G}}$  only. The orientation of this source is from the drain to the source.
- a source determined by  $K_{2g_{mb}}$ . This models the dependence of the drain current on  $v_{SB}$  only. The orientation of this source is from the source to the drain.
- a source determined by  $K_{2_{g_m\&g_{mb}}}$ . This models the dependence of the drain current of both  $v_{GS}$  and  $v_{SB}$ . The orientation of this source is from the drain to the source.

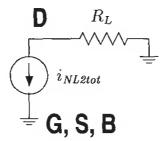


Figure 8.66: Linearized equivalent of the circuit of Figure 8.64 excited with the nonlinear current source of order two.

In the circuit of Figure 8.66 these three sources have been lumped into one single source  $i_{NL2tot}$ . The transconductances  $g_m$  and  $g_{mb}$  do not play a role in the circuit of Figure 8.66. The reason is that their AC controlling voltage is zero. Indeed, the gate and the bulk are at AC ground, while the source is fixed to the input voltage, which is assumed to be purely sinusoidal. As a result, no harmonics occur at the source node. Consequently, all controlling voltages in the linearized circuit that is excited by the nonlinear current sources of order two, are zero.

The nonlinear current source  $i_{NL2tot}$  is given by the sum of its three composing sources:

$$i_{NL2tot} = i_{NL2g_m} - i_{NL2g_{mb}} + i_{NL2g_m \& g_{mb}}$$
(8.344)

Using Table 5.5 we find for these sources

$$i_{NL2tot} = \frac{K_{2g_m}}{2} V_{gs,1,0}^2 - \frac{K_{2g_{mb}}}{2} V_{sb,1,0}^2 + \frac{K_{2g_m \& g_{mb}}}{2} V_{gs,1,0} V_{sb,1,0}$$
(8.345)

Since  $V_{gs,1,0} = -V_{in}$  and  $V_{sb,1,0} = V_{in}$  we find

$$i_{NL2tot} = \frac{1}{2} \left( K_{2g_m} - K_{2g_{mb}} - K_{2g_m \& g_{mb}} \right) \cdot V_{in}^2$$
 (8.346)

In Figure 8.66 it is seen that the current  $i_{NL2tot}$  is also the output current:

$$I_{out,2,0} = i_{NL2tot} \tag{8.347}$$

The second harmonic distortion is found as the ratio of the second- and first-order component of the output current:

$$HD_2 = \frac{V_{in}}{2(g_m + g_{mb})} \left[ K_{2g_m} - K_{2g_{mb}} - K_{2g_m \& g_{mb}} \right]$$
(8.348)

We now compare this expression to the expression of  $HD_2$  for a common-source amplifier. If in equation (8.348) we put  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{2g_{mb}\&g_{mb}}$  equal to zero then we obtain the same expression as for a common-source amplifier. For the common-gate transistor, extra terms occur

due to the bulk effect. Usually the nonlinearity coefficients  $K_{2g_{mb}}$  and  $K_{2g_{m}\&g_{mb}}$  are negative (see for example Table 8.4). On the other hand, the coefficient  $K_{2g_m}$  differs for the common-source and the common-gate configuration, due to the presence of the bulk effect in the latter case.

In order to make a fair comparison between a common-source and a common-gate stage, we evaluate  $HD_2$  of the two topologies by comparing two transistors with the same value of  $(V_{GS}-V_T)$ . For the common-gate transistor the bulk effect is present of course. For the commonsource transistor no bulk effect is present, but the gate-source voltage has been adjusted such that  $(V_{GS}-V_T)$  is the same as for the common-gate transistor. We already made such comparison in Section 8.5.2 using Table 8.4. With the data of this table we find for the common-gate transistor, using equation (8.348), a second-order intercept point  $IP_{2h}$  for harmonic distortion of 1.00V. For the common-source transistor we find an intercept point of 1.31V, which is higher. Hence, the common-source stage is more linear than the common-gate configuration.

Consider now the case of a bipolar common-base amplifier. Making  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{2g_{mb}}$  and  $K_{2g_{mb}}$ zero in equation (8.348) yields

$$HD_2 = K'_{2g_m} \cdot \frac{V_{in}}{2} \tag{8.349}$$

which is exactly the same as for a common-emitter amplifier. Here, of course, we neglected the base current.

#### 8.9.3 Third-order response

The third-order response is found by applying the nonlinear current sources of order three and finding the resulting output current. The nonlinear current sources of order three all appear in parallel. Their sum is given by

$$i_{NL3tot} = i_{NL3g_m} - i_{NL3g_{mb}} + i_{NL3g_m\&g_{mb}}$$
 (8.350)

The nonlinear current source  $i_{NL3q_m}$  is found from Table 5.7

$$i_{NL3gm} = \frac{K_{3g_m}}{4} V_{gs,1,0}^3 + K_{2g_m} V_{gs,2,0} V_{gs,1,0}$$
(8.351)

The second term of this current source is zero, since the second-order response  $V_{gs,2,0}$  is zero. Hence  $i_{NL3om}=-\frac{K_{3g_m}}{4}V_{in}^3 \tag{8.352}$ 

$$i_{NL3gm} = -\frac{K_{3g_m}}{4} V_{in}^3 \tag{8.352}$$

Similarly we find

$$i_{NL3g_{mb}} = \frac{K_{3g_{mb}}}{4} V_{in}^3 \tag{8.353}$$

and

$$i_{NL3}{}_{gm\&g_{mb}} = \frac{K_{3_{2g_m\&g_{mb}}}V_{gs,1,0}^2V_{sb,1,0} + \frac{K_{3_{g_m\&2g_{mb}}}V_{gs,1,0}V_{sb,1,0}^2}{4}V_{gs,1,0}V_{sb,1,0}^2 + \frac{K_{2g_m\&g_{mb}}}{2}\left(V_{gs,1,0}V_{sb,2,0} + V_{gs,2,0}V_{sb,1,0}\right)$$

$$= \frac{K_{3_{2g_m\&g_{mb}}}V_{in}^3 - \frac{K_{3g_m\&2g_{mb}}V_{in}^3}{4}V_{in}^3}{4}(8.354)$$

The third-order component  $I_{out,3,0}$  of the output current is equal to  $i_{NL3tot}$ . Combining the above results, we find the third harmonic distortion of the output current:

$$HD_{3} = \frac{V_{in}^{2}}{4(g_{m} + g_{mb})} \left( K_{3g_{m}} + K_{3g_{mb}} - K_{3g_{m} \& g_{mb}} + K_{3g_{m} \& 2g_{mb}} \right)$$
(8.355)

Again we can compare this expression to the third harmonic distortion of a common-source amplifier. With the two transistors of Table 8.4 we find a third-order intercept point  $IP_{3h}$  of 3.52V for the common-gate stage and a value of 4.4V for the common-source stage. Hence the common-source stage is again more linear.

From equation (8.355) one can easily derive  $HD_3$  for a common-base stage:

$$HD_3 = K'_{3g_m} \cdot \frac{V_{in}^2}{4} \tag{8.356}$$

which is exactly the same value as for a common-emitter stage.

# 8.10 Current mirrors

In this section we analyze the distortion generated in a simple current mirror. Since the analysis of bipolar and a MOS current mirrors is very similar, we discuss them in one section.

A bipolar and a MOS current mirror are shown in Figure 8.67 and Figure 8.68, respectively. A MOS current mirror has already been discussed qualitatively in Section 4.7.3 as an example of pre- and post-distortion.

We will first derive a DC transfer characteristic of a current mirror for a simple bipolar transistor model and find that the current mirror does not produce distortion when the simple transistor model holds and the two transistors of the current mirror match. The same is true when an elementary MOS transistor model is used.

Next, we will derive a symbolic expression for the second and third harmonic distortion in terms of nonlinearity coefficients and as a function of frequency. This expression will allow us to study the influence of capacitors and the influence of mismatches on distortion.

### 8.10.1 DC transfer characteristic for a bipolar current mirror

The operation of the bipolar current mirror of Figure 8.67 is as follows: the input current  $i_{IN}$  is transformed into a voltage by the diode-connected transistor  $Q_{1A}$ . This voltage controls the collector current of transistor  $Q_{1B}$ . This collector current is also the output current.

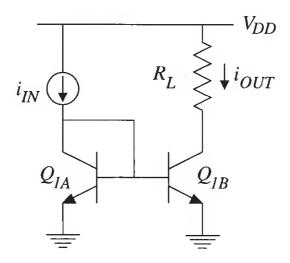


Figure 8.67: A bipolar current mirror.

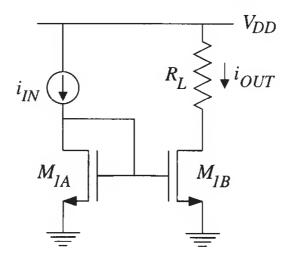


Figure 8.68: A MOS current mirror.

The current-to-voltage conversion by  $Q_{1A}$  is nonlinear. If  $Q_{1A}$  is modeled only with its collector current without Early effect, then under low-injection conditions the base-emitter voltage of transistor  $Q_{1A}$  is given by

$$v_{BE_{1A}} = V_t \ln \left( \frac{i_{IN}}{I_{S_{1A}}} \right) \tag{8.357}$$

in which  $I_{S_{1A}}$  is the saturation current of transistor  $Q_{1A}$ . The collector current of transistor  $Q_{1B_a}$  also the output current  $i_{OUT}$ . It is given by

$$i_{OUT} = i_{C_{1B}} = I_{S_{1B}} \exp\left(\frac{v_{BE_{1A}}}{V_t}\right)$$
 (8.358)

8.10 Current mirrors

Using equation (8.357) we find

$$i_{OUT} = i_{C_{1B}} = \frac{I_{S_{1B}}}{I_{S_{1A}}} i_{IN} \tag{8.359}$$

The ratio of the saturation currents depends on the ratio of the emitter areas of the two transistors. Equation (8.359) shows that the relationship between the output current and the input current is linear such that no distortion occurs. This is an example of pre- and post-distortion: the nonlinearity of the current-to-voltage conversion by transistor  $Q_{1A}$  is compensated by the nonlinear relationship between the base-emitter voltage and the collector current of transistor  $Q_{1B}$ . As a result, we have

$$HD_2 = 0$$
  $IM_2 = 0$  (8.360)  
 $HD_3 = 0$   $IM_3 = 0$  (8.361)

$$HD_3 = 0 IM_3 = 0 (8.361)$$

This is also true if a more complicated dependence of the collector current on the base-emitter voltage is considered. Further, the same conclusions hold for the MOS current mirror of Figure 8.68 when two transistors match and when only the dependence of the drain current on the gate-source voltage is taken into account.

The above analysis is only approximate. The influence of the base current, the output conductance, the ohmic resistors and the capacitors has been neglected. In Section 8.10.3 we will study the influence on distortion of the base current and the capacitor  $C_{\pi}$  of the two transistors  $Q_{1A}$ and  $Q_{1B}$ . Before studying this, we will first analyze the influence of mismatches and capacitors in a MOS current mirror, which yields simpler expressions than its bipolar counterpart.

#### Distortion in a MOS current mirror 8.10.2

Using the calculation method of Section 5.3, we will derive a closed-form expression for the second and third harmonic of the output current of the MOS current mirror of Figure 8.68. Afterwards, we will evaluate the results numerically in Section 8.10.2.4. The starting point for the computations is the nonlinear circuit of Figure 8.69 which is the AC-equivalent circuit of the one shown in Figure 8.68. In this equivalent circuit we take into account the nonlinear dependence

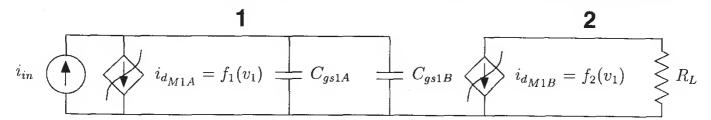


Figure 8.69: AC-equivalent circuit of the MOS current mirror of Figure 8.68.

of the drain current on the gate-source voltage of each transistor. The output conductance of the transistors is neglected. Further, the gate-source capacitance of each transistor has been taken into account. Finally, the transistors do not necessarily match.

### 8.10.2.1 First-order response

The response of the linearized circuit is computed first. This circuit is shown in Figure 8.70. In

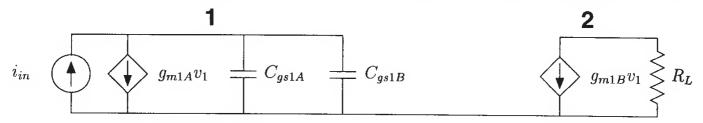


Figure 8.70: Linearized equivalent of the circuit of Figure 8.69.

this circuit we easily find the first-order response at node 1:

$$V_{1,1,0} = \frac{I_{in}}{g_{m1A} + sC_{gsTOT}} \tag{8.362}$$

in which  $C_{gsTOT}$  is the sum of the two gate-source capacitors.

The fundamental of the output current is given by the product of the transconductance of transistor  $M_{1B}$  with its gate-source voltage:

$$I_{out,1,0} = g_{m1B}V_{1,1,0} = \frac{g_{m1B}}{g_{m1A} + sC_{qsTOT}} \cdot I_{in}$$
(8.363)

### 8.10.2.2 Second-order response

For the computation of the second-order response, the linearized circuit of Figure 8.70 is excited with the nonlinear current sources of order two that correspond to the drain current nonlinearity of the two transistors. This situation is shown in Figure 8.71.

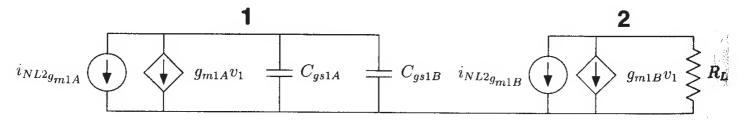


Figure 8.71: Linearized equivalent of the circuit of Figure 8.69 excited with nonlinear current sources of order two.

The value of the nonlinear current sources is found using Table 5.5:

$$i_{NL2g_{m1A}} = \frac{K_{2g_{m1A}}}{2} V_{1,1,0}^2 \tag{8.364}$$

$$i_{NL2g_{m1B}} = \frac{K_{2g_{m1B}}}{2} V_{1,1,0}^2 \tag{8.365}$$

The second-order response  $V_{1,2,0}$  at node 1 is seen to be

$$V_{1,2,0} = -\frac{i_{NL2g_{m1A}}}{g_{m1A} + 2sC_{gsTOT}}$$
(8.366)

Notice that the transfer function from the nonlinear current source to node 1 has been evaluated at 2s instead of s.

Using equation (8.364) we find for  $V_{1,2,0}$ 

$$V_{1,2,0} = -\frac{K_{2g_{m1A}}}{2} \frac{I_{in}^2}{(g_{m1A} + sC_{qsTOT})^2 (g_{m1A} + 2sC_{qsTOT})}$$
(8.367)

The second harmonic of the output current is the sum of the nonlinear current source that corresponds to  $K_{2g_{m1B}}$  and the current from the current generator  $g_{m1B}V_{1,2,0}$ :

$$I_{out,2,0} = i_{NL2g_{m1B}} + g_{m1B}V_{1,2,0} (8.368)$$

Using equations (8.365) and (8.367) and dividing the second-order response by the first-order one, we find for the second harmonic distortion

$$HD_{2} = \left| \frac{I_{in}}{2g_{m1B} \left( g_{m1A} + sC_{gsTOT} \right)} \cdot \left( K_{2g_{m1B}} - K_{2g_{m1A}} \frac{g_{m1B}}{g_{m1A} + 2sC_{gsTOT}} \right) \right|$$
(8.369)

It is seen that HD<sub>2</sub> at low frequencies is zero if

$$\frac{K_{2g_{m1B}}}{K_{2g_{m1A}}} = \frac{g_{m1B}}{g_{m1A}} \tag{8.370}$$

This condition is satisfied if the two transistors  $M_{1A}$  and  $M_{1B}$  are identical. Assume that the two transistors match perfectly but they have a different gate width. This occurs for example when one has to obtain a current gain higher than one. According to the model equation (7.121), condition (8.370) is satisfied since the current is linear with respect to W. However, this is not exact: effects such as the narrow-width effect (see Section 7.9) destroy this linear relationship. As a result, condition (8.370) is not satisfied.

Assume now that there is a mismatch in the threshold voltages of  $M_{1A}$  and  $M_{1B}$ . In Section 8.5.3 it was found that the relative variation on  $g_m$  due to the threshold voltage variation is different from the relative variation on  $K_{2g_m}$ . As a result, distortion will occur.

Finally, consider the influence of the gate-source capacitors. It is seen from equation (8.369) that distortion increases as the frequency of the AC input current increases. At high frequencies, a considerable amount of the input current flows into the linear gate-source capacitors. The presence of these linear capacitors disturbs the pre-distortion mechanism. It should be noted that the distortion at high frequencies caused by the gate-source capacitors is also present when the transistors  $M_{1A}$  and  $M_{1B}$  match perfectly.

The second harmonic distortion given by equation (8.369) will be evaluated numerically in Section 8.10.2.4.

The presence of a linear output conductance of the input current source also disturbs the predistortion mechanism just as the gate-source capacitors. The effect of this output conductance can be modeled starting from equation (8.369) by replacing  $sC_{gsTOT}$  with this output conductance.

### 8.10.2.3 Third-order response

For the computation of the third-order response, the nonlinear current sources of order three are applied to the linear circuit. This yields a similar situation as for order two, which is shown in Figure 8.71. The third-order nonlinear current sources are given by

$$i_{NL3g_{m1A}} = \frac{K_{3g_{m1A}}}{4} V_{1,1,0}^3 + K_{2g_{m1A}} V_{1,1,0} V_{1,2,0}$$
(8.371)

$$i_{NL3g_{m1B}} = \frac{K_{3g_{m1B}}}{4} V_{1,1,0}^3 + K_{2g_{m1B}} V_{1,1,0} V_{1,2,0}$$
(8.372)

The third-order response  $V_{1,3,0}$  at node 1 is found in a way similar to the computation for order two:

$$V_{1,3,0} = -\frac{i_{NL3g_{m1A}}}{g_{m1A} + 3sC_{gsTOT}}$$
(8.373)

The third harmonic of the output current is the sum of the current from the current generator  $g_{m1B}V_{1,3,0}$  and the nonlinear current source  $i_{NL3g_{m1B}}$ :

$$I_{out,3,0} = i_{NL3g_{m1B}} + g_{m1B}V_{1,3,0} (8.374)$$

After some algebra we find

$$HD_{3} = \left| \frac{I_{in}^{2}}{2g_{m1B} \left( g_{m1A} + sC_{gsTOT} \right)^{2}} \left( \frac{1}{2} \left( K_{3g_{m1B}} - \frac{g_{m1B}}{g_{m1A} + 3sC_{gsTOT}} K_{3g_{m1A}} \right) - \frac{K_{2g_{m1A}}}{g_{m1A} + 2sC_{gsTOT}} \left( K_{2g_{m1B}} - \frac{g_{m1B}}{g_{m1A} + 3sC_{gsTOT}} K_{2g_{m1A}} \right) \right) \right|$$
(8.375)

From this expression similar conclusions can be drawn as for order two. This expression is evaluated numerically in the next section.

From equation (8.375) we see that  $HD_3$  at low frequencies is zero if

$$\frac{K_{3g_{m1B}}}{K_{3g_{m1A}}} = \frac{g_{m1B}}{g_{m1A}} \quad \text{and} \quad \frac{K_{2g_{m1B}}}{K_{2g_{m1A}}} = \frac{g_{m1B}}{g_{m1A}}$$
(8.376)

This is satisfied when the transistors  $M_{1A}$  and  $M_{1B}$  are identical.

### 8.10.2.4 Numerical example

For the MOS current mirror of Figure 8.69 we now evaluate the second- and third-order intercept points  $IP_{2h}$  and  $IP_{3h}$  as a function of frequency and in the presence of a mismatch of the threshold voltage between the two transistors  $M_{1A}$  and  $M_{1B}$ . The intercept points are derived from equations (8.369) and (8.375). The dimensions of the two transistors are  $W = 80 \mu m$  and  $L = 0.7 \mu m$ . The bias voltages are  $V_{GS} = 1.25 V$  and  $V_{SB} = 0 V$ . The threshold voltage of

	$M_{1A}$	$M_{1B}$
$i_D$	0.865mA	0.899mA
$g_m$	3.37mA/V	3.43mA/V
$K_{2g_m}$	$3.04mA/V^2$	$3.02mA/V^{2}$
$K_{3g_m}$	$-0.618 mA/V^3$	$-0.624 mA/V^3$
$C_{gs}$	21 <i>fF</i>	21 <i>fF</i>

Table 8.5: Numerical values for the drain current and its derivatives for the transistors  $M_{1A}$  and  $M_{1B}$  in the MOS current mirror of Figure 8.68.

transistor  $M_{1A}$  is 750mV, which is the nominal value. In this way we find  $V_{GS} - V_T = 0.5V$ . The threshold voltage of transistor  $M_{1B}$  has been taken equal to 740mV. The drain current and its derivatives have been computed based upon the model equation (7.121). The results are given in Table 8.5.

The second- and third-order intercept points  $IP_{2h}$  and  $IP_{3h}$  are plotted as a function of frequency in Figure 8.72. The two intercept points decrease as the frequency increases, which means that the distortion increases with increasing frequency. At low frequencies, the limitation of the intercept point is caused by mismatches, whereas the main limitation at high frequencies is caused by the presence of the gate-source capacitors.

### 8.10.3 Distortion in a bipolar current mirror

In a bipolar current mirror four extra nonlinearities are present compared to the MOS current mirror of Figure 8.69, namely the two base currents and the two nonlinear capacitors  $C_{\pi}$ . In the same way as above one can compute the harmonics of the output current. The numerator of  $HD_2$  is given by

numerator of 
$$HD_2 = \left| -g_{m1A} K_{2g_{m1B}} + g_{m1B} K_{2g_{m1A}} - (g_{\pi 1A} + g_{\pi 1B}) K_{2g_{m1B}} + g_{m1B} K_{2g_{\pi 1B}} + g_{m1B} K_{2g_{\pi 1A}} + 2s \left( g_{m1B} K_{2C_{\pi 1B}} + g_{m1B} K_{2C_{\pi 1A}} - (C_{\pi 1A} + C_{\pi 1B}) K_{2g_{m1B}} \right) \right|$$
 (8.377)

and the denominator is given by

denominator of 
$$HD_2 = \left| 2g_{m1B} \left( g_{m1A} + g_{\pi 1A} + g_{\pi 1B} + s(C_{\pi 1A} + C_{\pi 1B}) \right) \right.$$

$$\left. \cdot \left( g_{m1A} + g_{\pi 1A} + g_{\pi 1B} + 2s(C_{\pi 1A} + C_{\pi 1B}) \right) \right|$$
(8.378)

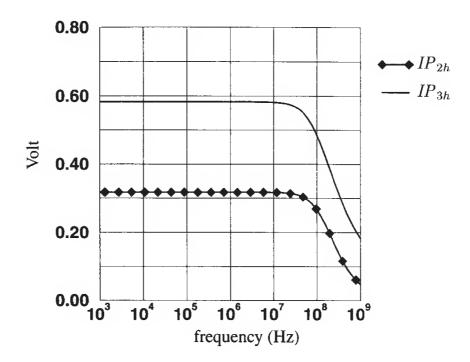


Figure 8.72: Second- and third-order intercept points  $IP_{2h}$  and  $IP_{3h}$  as a function of frequency for the MOS current mirror of Figure 8.68 with a mismatch between transistors  $M_{1A}$  and  $M_{1B}$ .

Assume now that the two transistors  $Q_{1A}$  and  $Q_{1B}$  match perfectly. In this case, the numerator of  $HD_2$  reduces to

numerator of 
$$HD_2 = \left(-2g_{\pi}K_{2g_m} + 2g_mK_{2g_{\pi}}\right) + 4s(g_mK_{2C_{\pi}} - C_{\pi}K_{2g_m})$$
 (8.379)

in which we dropped the subscripts "1A" and "1B" for convenience. It is seen that even with perfect matching  $HD_2$  can differ from zero. However, if we assume that the nonlinearities of the base current, collector current and  $C_{\pi}$  track, then it is easy to see that  $HD_2$  becomes zero. In this case, the presence of the base current nonlinearity and the nonlinear  $C_{\pi}$  do not disturb the predistortion operation. In practice, the nonlinearities will not track perfectly, such that a nonzero second harmonic distortion results. For the third harmonic distortion similar conclusions can be drawn.

# 8.11 Bipolar double-balanced mixer

A double-balanced mixer is often used in receivers and transmitters. The Gilbert multiplier with pre-distortion of Figure 4.21 can be used as a double-balanced mixer. A complete schematic of a bipolar double-balanced Gilbert mixer is shown in Figure 8.73.

Both in receive and transmit applications a mixer performs a frequency translation: in receiver a high-frequency signal is mixed with the signal of a local oscillator resulting in a low frequency signal and in a transmitter a low-frequency signal is translated to a high-frequency one. This translation can be performed by a weakly nonlinear mixer that makes use of a second-order

nonlinearity that generates from two input signals a second-order intermodulation product at the sum and at the difference frequency. Depending on the application, either the signal at the sum or the difference frequency is of interest. In this section, we analyze the conversion gain of the double-balanced mixer of Figure 8.73 for both receive and transmit applications.

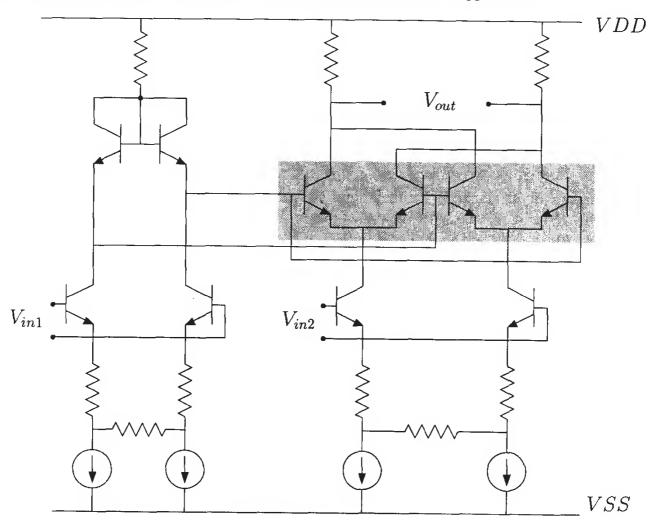


Figure 8.73: A bipolar double-balanced Gilbert mixer.

The conversion gain is a second-order intermodulation product. According to Table 5.5, this intermodulation product is given by

Output at 
$$|\omega_1 \pm \omega_2| = V_{LO}V_{in} \sum_i K_{2i} TF_{1i}(\omega_1) TF_{2i}(\omega_2) TF_{3i}(\omega_1 \pm \omega_2)$$
 (8.380)

in which  $V_{LO}$  is the amplitude of the local oscillator. The sum is taken over all basic nonlinearities in the circuit. In this expression it has been assumed for simplicity that all nonlinearities are one-dimensional conductances or transconductances. Other types of basic nonlinearities can be added similarly. The functions  $TF_{1i}(\omega_1)$  and  $TF_{2i}(\omega_2)$  are the transfer functions from the input and the local oscillator port respectively, to the voltage that determines the nonlinearity. The transfer function  $TF_3(\omega_1 \pm \omega_2)$  describes the transfer from the nonlinear current source of order two to

the output needs to be calculated. If the input amplitude  $V_{in}$  is set to 1 V for reference, then the conversion gain K is found from equation (8.380):

$$K = V_{LO} \sum_{i} K_{2i} TF_{1i}(\omega_1) TF_{2i}(\omega_2) TF_{3i}(\omega_1 \pm \omega_2)$$
 (8.381)

Applying this formula<sup>1</sup> to the double-balanced mixer of Figure 8.73, one finds that the conversion gain is primarily determined by the transconductance nonlinearity of the four transistors of the mixer core (enclosed in the shaded area in Figure 8.73). This is true both for transmit applications ( $V_{in}$  at low frequencies) and receive applications ( $V_{in}$  at high frequencies), and it is true regardless of the input port at which the input signal is applied ( $V_{in1}$  and  $V_{in2}$  in Figure 8.73). This means that in the sum of equation (8.381), only four terms are significant, namely the ones that correspond to  $K_{2g_m}$  of the mixer core transistors. Since the transistors in the mixer core ideally match, these terms are identical, except for their sign. Now the three transfer functions  $TF_1$ ,  $TF_2$  and  $TF_3$  can be analyzed in detail. In Figure 8.74 these transfer functions are shown as a function of frequency. Clearly, an overshoot occurs in the transfer functions  $TF_1$  and  $TF_2$ . This is due to

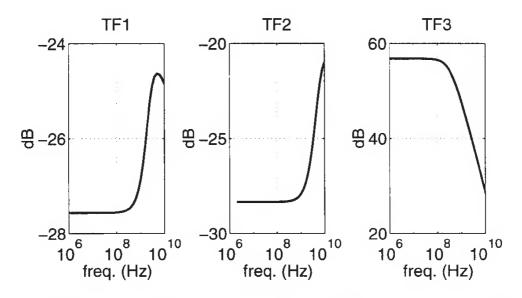


Figure 8.74: The transfer functions  $TF_1$ ,  $TF_2$  and  $TF_3$  as a function of frequency.

the inductive behavior of the common-base transistors with a base resistance [Gray 93, Lak 94]. When two high-frequency inputs at  $\omega_1$  (=2 $\pi f_1$ ) and  $\omega_2$  (=2 $\pi f_2$ ) are applied to the circuit used as a downconverter, then the transfer function  $TF_3$  must be evaluated at low frequencies, since  $|\omega_1 - \omega_2|$  is small. In this case, the overshoots of  $TF_1$  and  $TF_2$  result in an overshoot in the conversion gain at the difference frequency, which is shown in Figure 8.75a. When the mixer is used as an upconverter, then the output of interest is at high frequencies. The conversion gain at the sum frequency  $\omega_1 + \omega_2$  is shown in Figure 8.75b. It is seen that no overshoot occurs since the transfer function  $TF_3$  now must be evaluated at high frequencies, where its response is not flat anymore.

<sup>&</sup>lt;sup>1</sup>Basic nonlinearities other than one-dimensional (trans)conductances have been taken into account as well.

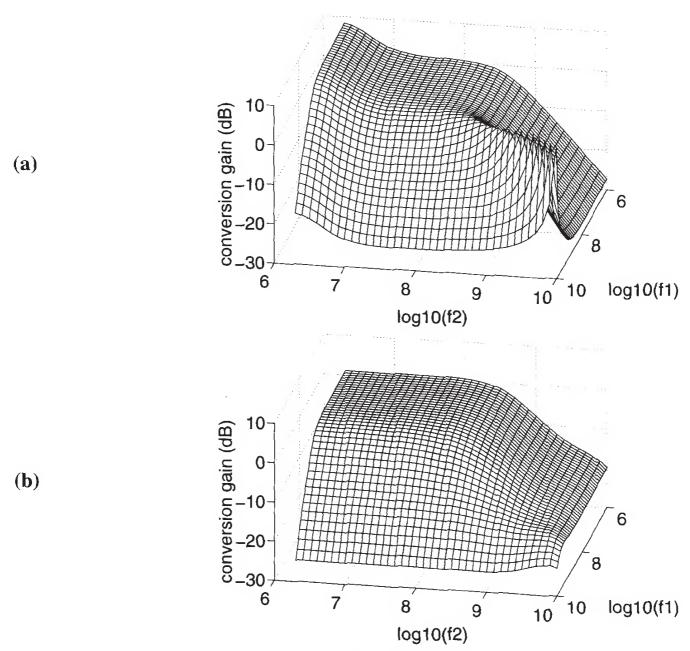


Figure 8.75: Conversion gain at the difference frequency (a) and at the sum frequency (b) of the double-balanced mixer as a function of the two input frequencies. The value is scaled to an amplitude of IV for both the signal input and the local oscillator input.

# 8.12 CMOS Miller-compensated operational amplifier

In this example we analyze the second harmonic at the output of the CMOS two-stage Miller-compensated operational amplifier of Figure 5.11 up to its gain-bandwidth product (GBW). The amplifier is put in the inverting feedback configuration of Figure 4.30. It is assumed that the transistors  $M_{1A}$  and  $M_{1B}$  match, just as  $M_{2A}$  and  $M_{2B}$ . Small-signal parameters and nonlinearity coefficients of such matching transistors are represented with one symbol. For example,  $g_{01}$  represents the output conductance of both  $M_{1A}$  and  $M_{1B}$ .

The amplifier has a gain-bandwidth of  $100 \, kHz$ , the load capacitance  $C_L$  is 10 pF, the load resistance  $R_L$  in addition to the load formed by the feedback resistors is  $100 k\Omega$  and the compensation capacitance equals 1 pF.

The second harmonic distortion  $HD_2$  has sixteen contributions. These are first calculated as a function of the fundamental frequency with ISAAC using the method explained in Section 5.3. The most important contributions to  $HD_2$  are shown in Figure 8.76. The other contributions are below  $-70\,dB$ .

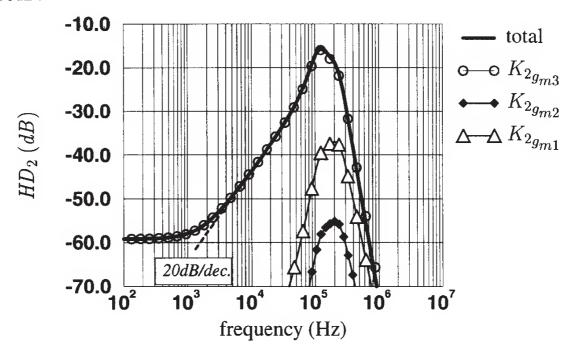


Figure 8.76: The most important contributions to the second harmonic distortion of the output voltage of the CMOS two-stage Miller-compensated opamp of Figure 5.11 in the feedback configuration of Figure 4.30.

It is seen that  $HD_2$  starts to increase from 1 kHz ( $\approx 0.01 GBW$ ) with 20 dB per decade, and from about 50 kHz ( $\approx GBW/2$ )with a steeper slope. Beyond the gain-bandwidth  $HD_2$  decreases rapidly. This behavior is also seen in measurements, as shown in Figure 8.77. For the third harmonic distortion, a similar increase with frequency is seen, but with a steeper slope. Differences in absolute value between the computed and measured distortion levels are mainly due to the poor modeling of the output conductance with the available SPICE level 2 models.

Clearly, only one nonlinearity dominates for frequencies below the gain-bandwidth produc GBW (100kHz), namely the second-order nonlinearity coefficient  $K_{2g_m}$  of transistor  $M_3$ . The can be explained by the fact that the largest contributions to the nonlinear distortion at the output of an amplifier originate from the circuit elements close to, or at the output, where signal swin are large.

An expression for  $HD_2$  can be computed by considering the contribution of  $K_{2g_{m3}}$  only. This first computed in open loop. It is given by

contribution of 
$$K_{2g_{m3}} = i_{NL2g_{m3}} \cdot TF_{i_{NL2g_{m3}} \rightarrow out}(2\omega_1)$$
 (8.38)

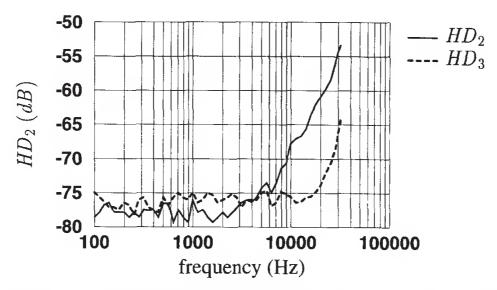


Figure 8.77: Measured second and third harmonic distortion on the CMOS two-stage Miller-compensated opamp of Figure 5.11 in the feedback configuration of Figure 4.30.

From Table 5.5 the value of  $i_{NL2g_{m3}}$  is found:

$$i_{NL2g_{m3}} = \frac{K_{2g_{m3}}}{2} V_{gs3,1,0}^2 \tag{8.383}$$

in which  $V_{gs3,1,0}$  is the fundamental response of the gate-source voltage of  $M_3$ . At frequencies well below GBW this is easily found to be

$$V_{gs3,1,0} \approx -\frac{g_{m1} \left(g_L + j\omega_1 \left(C_L + C_C\right)\right)}{g_{o1} + g_{o2} + j\omega_1 C_C \left(g_{m3}/g_L\right)} V_{in}$$
(8.384)

in which  $g_L$  is the sum of  $1/R_L$  and the output conductances of  $M_3$  and  $M_4$ .

The nonlinear current source of order two that corresponds to  $K_{2g_{m3}}$  flows from the drain of  $M_3$  to its source. The transfer function from this source to the output of the amplifier, which is the drain of  $M_3$ , is given by

$$TF_{i_{NL2}g_{m3} \to out}(2\omega_1) = -\frac{g_{o1} + g_{o2} + 2j\omega_1 C_C}{g_L(g_{o1} + g_{o2} + 2j\omega_1(g_{m3}/g_L))}$$
(8.385)

In order to know the second harmonic in closed loop, the second harmonic in open loop needs to be divided by  $[(1 + T(j\omega_1))^2 (1 + T(2j\omega_1))]$ , T being the loop gain as explained in Section 4.8. The second harmonic distortion is obtained by dividing the second harmonic by the fundamental response. Doing so, the poles in equations (8.384) and (8.385) are canceled, and  $HD_2$  is computed with the above method as

$$HD_{2} = \frac{1}{2} V_{in} \frac{(R_{1} + R_{2}) R_{2}}{R_{1}^{2}} \frac{|(g_{L} + j\omega_{1}(C_{L} + C_{C}))^{2} (g_{o1} + g_{o2} + 2j\omega_{1}C_{C})|}{g_{m1}g_{m3}^{3}} K_{2g_{m3}}$$
(8.386)

It is seen that  $HD_2$  increases with 20 dB per decade from the frequency  $(g_{o1} + g_{o2})/(4\pi C_C)$ . For the given design this frequency is computed to be 2.3kHz. At the frequency  $g_L/(2\pi(C_L + C_C))$ , which equals 31kHz here, the increase is with 60 dB per decade.

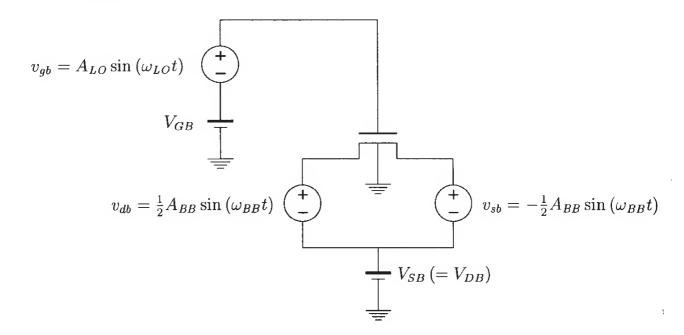


Figure 8.78: Principle schematic of a single-MOS-transistor mixer used as an upconverter. The transistor is biased in the triode region with  $v_{DS} = 0V$ .

# 8.13 CMOS upconverter

In Section 7.4.1.1 we already explained the principle of a single-MOS-transistor mixer. The MOS transistor is biased in the triode region with  $v_{DS} = 0V$ . The inputs to the mixing transistor are a voltage applied between the gate and the bulk, and a differential signal applied between source and drain. When the two input signals are sinusoidal, then the output of interest is the component in the drain current at the sum or difference frequency of the two input frequencies.

We will now analyze the single-transistor mixer which behaves in a weakly nonlinear way. Hereby we will use the MOS models of Section 7.7. We will concentrate on an upconverter. This means that a high-frequency local oscillator signal at  $\omega_{LO}$  is mixed with a low-frequency signal at a frequency  $\omega_{BB}$  and the output of interest is at the frequency  $\omega_{LO} + \omega_{BB}$  or at  $\omega_{LO} - \omega_{BB}$ .

A principle schematic is shown in Figure 8.78. The high-frequency local oscillator signal applied between the gate and the bulk. It is given by

$$v_{ab} = v_{LO}(t) = A_{LO}\sin(\omega_{LO}t) \tag{8.38}$$

The baseband signal is applied between drain and source. It is given by

$$v_{ds} = v_{BB}(t) = A_{BB}\sin(\omega_{BB}t) \tag{8.38}$$

Equivalently, we can say that half of the baseband signal is applied between source and bulk, at the other half between drain and bulk:

$$v_{db} = \frac{1}{2} A_{BB} \sin\left(\omega_{BB} t\right) \tag{8.38}$$

$$v_{sb} = -\frac{1}{2}A_{BB}\sin\left(\omega_{BB}t\right) \tag{8.3}$$

Several realizations of this principle have been reported [King 97, Borre 97]. The upconverter described in [King 97] will be analyzed more in detail in this section. It is shown in Figure 8.79. This is a 1 *GHz* upconversion mixer realized in the same CMOS technology that has been used in Chapter 7 for the computations of nonlinearity coefficients. The most important SPICE level 2 and 3 model parameters of this technology are shown in Table 7.1.

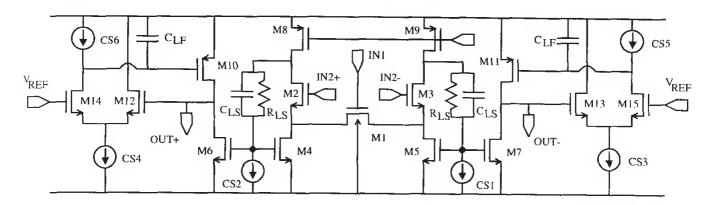


Figure 8.79: Schematic of a 1 GHz CMOS upconversion mixer [King 97], based upon the principle schematic of Figure 8.78.

The operation of the circuit of Figure 8.79 is as follows: the local oscillator signal is applied at the IN1 input and the differential baseband signal at the IN2+ and the IN2-) input. The source followers  $M_2$  and  $M_3$  copy the signal at their gate to their source. In this way the differential baseband signal is applied at the drain and source terminals of the mixing transistor  $M_1$  which operates in the triode region with  $v_{DS}=0V$ . As mentioned above, the output of interest is the component at the sum frequency of the AC drain current of  $M_1$ . This current is fed into the current buffer that consists of the transistors  $M_3$ ,  $M_5$  and  $M_7$  and similarly into the buffer consisting of the transistors  $M_2$ ,  $M_4$  and  $M_6$ . These current buffers transfer the current to the output. By connecting a resistive load between OUT+ and OUT- the output current is transformed into a differential output voltage.

The current buffers used in this circuit only have n-MOS transistors in their signal path. In this way, they can operate up to very high frequencies, in first order up to half of the cutoff frequency. In addition, the internal feedback of the current buffers lowers the input impedance, which is desirable for a current buffer. The input voltage range for the baseband signal at IN2 is extended by biasing the drains of the cascode transistors  $M_2$  and  $M_3$  at a higher voltage. This is achieved with a high ohmic resistor  $R_{LS}$  and a current source that controls the current through and thus the voltage across the resistor. The maximum input voltage range is extended by the voltage drop across the resistor. A capacitor  $C_{LS}$  assures the feedback at high frequencies and forms a capacitive divider with the gate capacitance of  $M_4$  and  $M_6$ .

In the circuit of Figure 8.79 the baseband signal that is present in the output current of the mixing transistor  $M_1$  is also copied to the output. This unwanted signal is suppressed by an active coil circuit at the output. The active coil circuit consists of the transistors  $M_{10}$ ,  $M_{12}$  and  $M_{14}$ . The differential pair and the feedback assure that the output voltage (at the drain of transistor  $M_{10}$ ) is

kept at  $V_{REF}$ . When a low-frequency current signal is applied at the drain of  $M_{10}$  then the shunt feedback limits the voltage swing at the drain of  $M_{10}$  such that a low impedance is obtained at that node. However, a higher frequencies the loop gain of this feedback drops through the effect of the capacitor  $C_{LF}$ . Hence, the input impedance increases linearly with frequency like in a passive coil.

The upconverter of Figure 8.79 has been processed and measured. The measured output spectrum is shown in Figure 8.80. The baseband signal is a 20kHz differential sinusoidal voltage

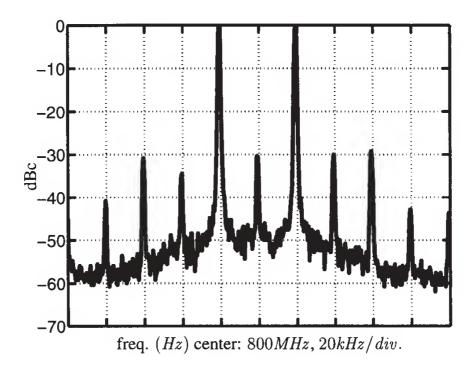


Figure 8.80: Measured output spectrum of the upconverter of Figure 8.79. The baseband signal is a 20kHz differential sinusoidal voltage with an amplitude of 1V. The LO signal is a 0dBm sinewave at 0.8GHz.

with an amplitude of 1V. The LO signal is a 0dBm sinewave at 0.8GHz.

The output spectrum is measured with an external balun (balanced-to-unbalanced RF tran former). The LO-feedthrough is at -30dBc. This feedthrough would not occur in a perfect balanced circuit. Also, the drain current of the mixing transistor does not contain a companent at  $\omega_{LO}$ , since at  $V_{DS}=0V$  the transconductance  $g_m$  or  $g_{mg}$  of  $M_1$  is zero. Hence LO-feedthrough is caused by imbalances and mismatches in the circuit at high frequencies. Sitilarly, the response at  $\omega_{LO}\pm 2\omega_{BB}$  is due to imbalances and mismatches only, as will be evidence below. The largest parasitic component occurs at  $\omega_{LO}\pm 3\omega_{BB}$ . The magnitude of this component will now be explained by making use of the MOS model of equation (7.105) that has been us in Section 7.7.4 to derive expressions for nonlinearity coefficients.

In order to find the values of the different spectral components of Figure 8.80 we limit analysis to the mixing transistor only. Then we can use the simplified circuit of Figure 8.5. This circuit is so simple that we do not explicitly need the method of Section 5.3 to compute

different intermodulation products of interest. Instead, we can limit ourselves to an analysis of a power series expansion of the drain current as a function of the different terminal voltages. We will express the current as a function of voltages referred to the bulk. In Chapter 7 such series expansions have been given up to order three. In this case, however, we will need fourth-order terms as well in order to explain the value of the spectral component at  $\omega_{LO} \pm 3\omega_{BB}$ . The power series expansion is a function of the AC values of the terminal voltages of the mixing transistor. The AC voltage  $v_{gb}$  corresponds to the LO input signal (see equation (8.387)) while  $v_{db}$  and  $v_{sb}$  are proportional to the baseband signal (see equations (8.389) and (8.390)):

$$\begin{split} i_{d} &= & g_{mg} \cdot v_{gb} + K_{2g_{mg}} \cdot v_{gb}^{2} + K_{3g_{mg}} \cdot v_{gb}^{3} + g_{md} \cdot v_{db} + K_{2g_{md}} \cdot v_{db}^{2} + K_{3g_{md}} \cdot v_{db}^{3} \\ & - g_{ms} \cdot v_{sb} - K_{2g_{ms}} \cdot v_{sb}^{2} - K_{3g_{ms}} \cdot v_{sb}^{3} \\ & + K_{2g_{mg} \& g_{ms}} \cdot v_{gb} \cdot v_{sb} + K_{32g_{mg} \& g_{ms}} \cdot v_{gb}^{2} \cdot v_{sb} + K_{3g_{mg} \& 2g_{ms}} \cdot v_{gb} \cdot v_{sb}^{2} \\ & + K_{2g_{mg} \& g_{md}} \cdot v_{gb} \cdot v_{db} + K_{32g_{mg} \& g_{md}} \cdot v_{gb}^{2} \cdot v_{db} + K_{3g_{mg} \& 2g_{md}} \cdot v_{gb} \cdot v_{db}^{2} \\ & + K_{2g_{ms} \& g_{md}} \cdot v_{sb} \cdot v_{db} + K_{32g_{ms} \& g_{md}} \cdot v_{sb}^{2} \cdot v_{db} + K_{3g_{ms} \& 2g_{md}} \cdot v_{sb}^{2} \cdot v_{db}^{2} \\ & + K_{3g_{mg} \& g_{ms} \& g_{md}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db} \\ & - K_{4g_{ms}} \cdot v_{sb}^{4} + K_{43g_{ms} \& g_{md}} \cdot v_{sb}^{3} \cdot v_{db} + K_{42g_{ms} \& 2g_{md}} \cdot v_{sb}^{2} \cdot v_{db}^{2} \\ & + K_{4g_{ms} \& 3g_{md}} \cdot v_{sb} \cdot v_{db}^{3} + K_{4g_{md}} \cdot v_{db}^{4} + K_{4g_{mg} \& 3g_{ms}} \cdot v_{sb}^{3} \cdot v_{gb} \\ & + K_{42g_{mg} \& 2g_{ms}} \cdot v_{sb}^{2} \cdot v_{gb}^{2} + K_{43g_{mg} \& 2g_{md}} \cdot v_{db}^{2} \cdot v_{gb}^{2} + K_{4g_{mg} \& 3g_{md}} \cdot v_{db}^{3} \cdot v_{gb} \\ & + K_{4g_{mg} \& g_{ms} \& 2g_{md}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db}^{2} + K_{4g_{mg} \& 2g_{ms} \& g_{md}} \cdot v_{gb}^{3} \cdot v_{sb}^{3} \cdot v_{db} \\ & + K_{4g_{mg} \& g_{ms} \& 2g_{md}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db}^{2} + K_{4g_{mg} \& 2g_{ms} \& g_{md}} \cdot v_{gb} \cdot v_{sb}^{2} \cdot v_{db} \\ & + K_{4g_{mg} \& g_{ms} \& 2g_{md}} \cdot v_{gb}^{2} \cdot v_{sb} \cdot v_{db} \end{split}$$

In Figure 8.80 we only see spectral components of the form  $\omega_{LO} \pm n\omega_{BB}$  with  $n=0,1,2,3,\ldots$ . Other spectral components fall outside the frequency band of interest. The responses seen in Figure 8.80 are linear with respect to the LO signal. Hence, we can limit ourselves to an analysis of the terms that are linear in  $v_{gb}$  in the power series expansion of equation (8.391).

**First-order term** The first-order term in equation (8.391) that is proportional to  $v_{gb}$  is given by

first-order term = 
$$g_{mg} \cdot v_{gb}$$
 (8.392)

From Table 7.9 we find that  $g_{mg}$  is zero for  $V_{DS}=0V$ . This means that the drain current does not contain a component at  $\omega_{LO}$ . The component at  $\omega_{LO}$  seen in Figure 8.80 is due to an imbalance in the circuit at high frequencies, such that the LO-signal feeds through to the output.

**Second-order terms** The second-order terms in equation (8.391) that are proportional to  $v_{gb}$  are given by

$$second-order\ terms = K_{2_{g_{mg}\&g_{ms}}} \cdot v_{gb} \cdot v_{sb} \ + \ K_{2_{g_{mg}\&g_{md}}} \cdot v_{gb} \cdot v_{db} \tag{8.393}$$

At  $V_{DS}=0V$  the source and the drain terminal play an identical role. In Section 7.4.2 it has been pointed out that in this case the derivatives of  $i_D$  with respect to  $v_{DB}$  are opposite to the derivatives with respect to  $v_{SB}$ . This has been shown only for a large transistor without taking into account effects such as mobility reduction and velocity saturation. However, even with the inclusion of these effects the derivatives are still opposite for  $v_{DB}=v_{SB}$ . Since  $v_{sb}=-v_{db}$  it is clear that the two terms in the right-hand side of equation (8.393) are identical. Equation (8.393) then becomes

second-order terms = 
$$2K_{2q_{mg}\&q_{md}} \cdot v_{gb} \cdot v_{db}$$
 (8.394)

Substituting now the expressions for  $v_{gb}$  and  $v_{db}$ , equations (8.387) and (8.389), and using some trigonometry (see Appendix A) yields the amplitude of the drain current components at the sum and the difference frequency:

|drain current component at 
$$(\omega_{LO} \pm \omega_{BB})$$
|  $= \frac{1}{2} \cdot A_{LO} \cdot A_{BB} \cdot K_{2g_{mg} \& g_{md}}$  (8.395)

An approximate expression for  $K_{2g_{mg}\&g_{md}}$  can be obtained from Table 7.9. Hereby, it has been checked that the accuracy is maintained on this expression, which is used here at different bias conditions than the ones at which the approximation has been made in Table 7.9. Using this expression we obtain

$$|\text{drain current component at } (\omega_{LO} \pm \omega_{BB})| = \frac{1}{2} \cdot A_{LO} \cdot A_{BB} \cdot \mu_0 C'_{ox} \frac{W}{L} \cdot \frac{1}{1 + \theta f_{\mu}}$$
(8.396)

In order to find the conversion gain, equation (8.395) must be multiplied with the loss of the source followers, the loss of the current buffers and with the load resistance at the output of the complete circuit. It is seen that the only design parameter with which the value of the conversion gain can be controlled, is the ratio W/L. Further it is seen that the conversion gain is degraded by mobility reduction.

**Third-order terms** The third-order terms in equation (8.391) that are proportional proportional to  $v_{ab}$  are given by

third-order terms = 
$$K_{3g_{mg}\&2g_{ms}} \cdot v_{gb} \cdot v_{sb}^2 + K_{3g_{mg}\&2g_{md}} \cdot v_{gb} \cdot v_{db}^2 + K_{3g_{mg}\&g_{ms}\&g_{md}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db}$$
 (8.397)

According to Table 7.3, the first two nonlinearity coefficients in this equation are proportional to the following derivatives:

$$K_{3_{g_{mg}\&2g_{ms}}} = \frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB} \partial v_{SB}^2} \tag{8.398}$$

and

$$K_{3g_{mg}\&2g_{md}} = \frac{1}{2} \frac{\partial^3 i_D}{\partial v_{GB} \partial v_{DB}^2}$$
(8.399)

For  $v_{DB} = v_{SB}$  these derivatives are opposite. Hence,  $K_{3g_{mg}\&2g_{ms}} = K_{3g_{mg}\&2g_{md}}$  at  $v_{DS} = 0V$ . The third nonlinearity coefficient in equation (8.397) is given by

$$K_{3_{g_{mg}\&g_{ms}\&g_{md}}} = \frac{\partial^{3}i_{D}}{\partial v_{GB}\partial v_{SB}\partial v_{DB}}$$
(8.400)

which is zero at  $v_{DS} = 0V^2$ .

As a result, the third-order component in the drain current that is proportional to  $v_{gb}$  is zero. The component observed in the measurement result of Figure 8.80 is due to imbalances and feedthrough at high frequencies.

Fourth-order terms in equation (8.391) that are proportional to  $v_{gb}$  are given by

$$\begin{aligned} \text{fourth-order terms} = & K_{4_{g_{mg}\&3g_{ms}}} \cdot v_{sb}^3 \cdot v_{gb} + K_{4_{g_{mg}\&3g_{md}}} \cdot v_{db}^3 \cdot v_{gb} \\ & + K_{4_{g_{mg}\&2g_{ms}\&2g_{md}}} \cdot v_{gb} \cdot v_{sb} \cdot v_{db}^2 + K_{4_{g_{mg}\&2g_{ms}\&g_{md}}} \cdot v_{gb} \cdot v_{sb}^2 \cdot v_{db} \end{aligned} \tag{8.401}$$

The fourth-order coefficients are proportional to the following derivatives:

$$K_{4_{g_{mg}\&3g_{ms}}} = \frac{1}{3!} \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB}^3}$$
 (8.402)

$$K_{4_{g_{mg} \& 3g_{md}}} = \frac{1}{3!} \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{DB}^3}$$
 (8.403)

$$K_{4_{g_{mg} \& g_{ms} \& 2g_{md}}} = \frac{1}{2!} \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB} \partial v_{DB}^2}$$
(8.404)

$$K_{4_{g_{mg}\&2g_{ms}\&g_{md}}} = \frac{1}{2!} \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB}^2 \partial v_{DB}}$$
(8.405)

Using the symmetry of source and drain at  $v_{DS}=0V$  we find that

$$\frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB}^3} = -\frac{\partial^4 i_D}{\partial v_{GB} \partial v_{DB}^3} \tag{8.406}$$

$$\frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB} \partial v_{DB}^2} = -\frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB}^2 \partial v_{DB}}$$
(8.407)

Further we have  $v_{sb} = -v_{db}$ . Hence we can combine the four terms in equation (8.401) to one single term:

fourth-order terms = 
$$K_4 v_{ab} v_{sb}^3$$
 (8.408)

<sup>&</sup>lt;sup>2</sup>This is not clear from the expression of  $K_{3g_{mg}\&g_{ms}\&g_{md}}$  in Table 7.9: the approximate expression in this Table has been derived for a value of  $v_{DS}$  different from zero.

in which the coefficient  $K_4$  is given by

$$K_4 = \frac{2}{3!} \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB}^3} + \frac{\partial^4 i_D}{\partial v_{GB} \partial v_{SB} \partial v_{DB}^2}$$
(8.409)

The AC voltages  $v_{gb}$  and  $v_{sb}$  are sinusoidal voltages, given by the equations (8.387) and (8.390), respectively. Using trigonometry (see Appendix A) we find the response at  $(\omega_{LO} \pm 3\omega_{BB})$  from equation (8.408):

|drain current component at 
$$(\omega_{LO} \pm 3\omega_{BB})$$
| =  $A_{LO} \cdot A_{BB^3} \cdot \frac{K_4}{64}$  (8.410)

Using the routines described in Section 3.5 an approximation for the coefficient  $K_4$  has been computed. With these routines it is seen that the largest term of  $K_4$  is 62 times larger than the second largest term. Hence we can approximate  $K_4$  by this one term:

$$K_4 \approx -\mu_0 C'_{ox} \frac{W}{L} \frac{2}{L^2 E_c^2 (1 + \theta f_\mu)}$$
 (8.411)

with  $f_{\mu}$  given by equation (7.68). The error on  $K_4$  that is made by taking into account the largest term only is 2.8%.

The factor  $L^2E_c^2$  arises from taking the derivative of the function hot, defined in equation (7.104), which describes velocity saturation. This means that the fourth-order response is almost completely determined by the variation of the drain current due to velocity saturation. Indeed, when the instantaneous drain current becomes high during operation, then the velocity of the carriers may saturate.

It is not a surprise that velocity saturation is by far the dominant factor that determines the fourth-order response. This response is proportional to derivatives in which three times a derivative with respect to  $v_{SB}$  (or  $v_{DB}$ ) has been taken, as seen in equation (8.409). In Figure 7.18 a strong influence of velocity saturation was already noticed on  $K_{3g_o}$ , which is proportional to the third-order derivative of the drain current with respect to the drain-source voltage.

The ratio of the fourth-order response at  $(\omega_{LO} \pm 3\omega_{BB})$  and the second-order response can be found by combining equations (8.396), (8.410) and (8.411). This yields

$$\left| \frac{\text{drain current component at } (\omega_{LO} \pm 3\omega_{BB})}{\text{drain current component at } (\omega_{LO} \pm \omega_{BB})} \right| = \frac{A_{BB}^2}{16L^2E_c^2}$$
 (8.412)

We will now evaluate this ratio for the mixer transistor in the circuit of Figure 8.79. The mixer transistor has a gate length of  $0.7\mu m$ . The critical field is given by

$$E_c = \frac{v_{sat}}{\mu_{eff}} \tag{8.41}$$

From the data of the  $0.7\mu m$  process of Table 7.1 we know that  $\mu_0$  is  $0.047m^2/(V.s)$ . The mobilireduction factor  $f_\mu$  is 0.146 in this case. With these data we find that the effective mobility  $\mu$  is  $0.041m^2/(V.s)$  and  $E_c=2.439MV/m$ . When the input amplitude  $A_{BB}$  equals 1V as in the second content of the secon

measurements of Figure 8.80 we find that the ratio given in equation (8.412) is -33dB, which is only 3dB different from the measured ratio.

The ratio between the fourth-order signal and the wanted signal of equation (8.412) can also be derived in another more intuitive way, using the knowledge that velocity saturation is the dominant effect that determines the fourth-order response. The drain current in the triode region is given by equation (7.92), which is repeated here for convenience:

$$i_D = large(v_{GB}, v_{DB}, v_{SB}) \cdot mobred(v_{GB}, v_{DB}, v_{SB}) \cdot hot(v_{GB}, v_{DB}, v_{SB})$$

$$(8.414)$$

with the functions *large*, *mobred* and *hot* given in equations (7.90), (7.73) and (7.104), respectively. The function *hot* can be approximated as follows:

$$hot(v_{GB}, v_{DB}, v_{SB}) = \frac{1}{\sqrt{1 + \left(\frac{v_{DB} - v_{SB}}{LE_c}\right)^2}} \approx 1 - \frac{1}{2} \left(\frac{v_{DB} - v_{SB}}{LE_c}\right)^2$$
(8.415)

Using this approximation in the drain current expression and neglecting the  $\frac{3}{2}$  powers in the function large, we can rewrite the drain current as

$$i_{D} \approx \mu_{0} C'_{ox} \frac{W}{L} \left[ 1 - \frac{1}{2} \left( \frac{v_{DB} - v_{SB}}{LE_{c}} \right)^{2} \right] \cdot \frac{1}{1 + \theta f_{\mu}} \cdot \left[ \left( v_{GB} - V_{FB} - \phi - \gamma \sqrt{\phi + v_{SB}} \right) (v_{DB} - v_{SB}) - \frac{1}{2} (v_{DB} - v_{SB}) (v_{DB} + v_{SB}) \right]$$
(8.416)

The baseband signal is applied over the drain and source terminals with  $V_{DS}=0V$ , and the local oscillator signal is applied at the gate. The AC drain current has two components that are proportional to  $v_{ab}$ :

$$i_{d1} = \mu_0 C'_{ox} \cdot \frac{W}{L} \cdot \frac{v_{gb} v_{ds}}{1 + \theta f_{\mu}}$$

$$\tag{8.417}$$

$$i_{d2} = \mu_0 C'_{ox} \cdot \frac{W}{L} \cdot \frac{1}{2} \cdot \left(\frac{v_{ds}}{LE_c}\right)^2 \cdot \frac{v_{gb}v_{ds}}{1 + \theta f_{\mu}}$$

$$(8.418)$$

For the computation of the wanted signal, velocity saturation is a higher-order effect, such that the wanted signal is derived from  $i_{d1}$ . The value of the fourth-order signal, on the other hand, can be found from  $i_{d2}$ . Substituting the values of  $v_{gb}$  and  $v_{ds}$  given in equations (8.387) and (8.388) into  $i_{d1}$  and  $i_{d2}$ , respectively, and using some trigonometry, one will find that the ratio of the fourth-order signal and the wanted second-order signal is as given in equation (8.412).

# 8.14 Summary

In this chapter we have studied the weakly nonlinear behavior of several basic MOS and bipolar circuits. To this purpose we have interpreted expressions that have been generated using the

calculation method of Section 5.3. This method is implemented in the symbolic network analysis program ISAAC that has been used throughout this chapter.

In the calculations we have taken into account effects that are difficult to take into account with hand calculations using Taylor series, such as frequency dependence of the nonlinear behavior, mismatches and the presence of parasitic circuit elements that make the derivation of a closed-form expression for the input-output characteristic of a circuit impossible.

A practical circuit usually contains many basic nonlinearities that give a contribution to the second or third harmonic at the circuit output. In order to obtain interpretable results, we have used several times the following procedure explained in Section 5.4.2.1: first the dominant contributions are determined, next an expression is generated for the dominant contributions.

In the example circuits in this chapter distortion is often suppressed with the techniques that have been studied in general in Chapter 4: pre-distortion, suppression of even harmonics by balanced operation, suppression of nonlinear behavior by linear feedback, . . . . In practice these principles are less effective due to parasitic effects such as mismatches or nonlinearities in the feedback path. These effects have also been considered in the example circuits.

It is seen throughout the examples in this chapter that the conversion of a voltage to a current is more difficult to realize in a linear way than a current-to-current conversion or a voltage-to-voltage conversion. For the latter two conversions pre- and post-distortion techniques can be used, as illustrated with the current mirrors.

# **Chapter 9**

# Measurements of basic nonlinearities of transistors

### 9.1 Introduction

In the absence of reliable transistor models and/or an accurate parameter extraction, one could consider to measure nonlinearity coefficients. This chapter concentrates on how higher-order derivatives of the transistor current can be measured accurately. The measurement results could be used in device parameter extraction. Only measurements on a bipolar transistor are presented in this chapter. The principles can be applied to MOS transistors as well.

The idea behind the measurements, explained in Section 9.2, is that harmonics and intermodulation products are measured that are determined by only one nonlinearity coefficient of one basic nonlinearity of the transistor. In this way, the value of this coefficient can be extracted from the measured harmonic or intermodulation product using expressions (2.13), (2.14), (2.29) and (2.30).

With the measurement method, values have been obtained for the nine different coefficients that describe the nonlinear dependence of the collector current on the base-emitter and the collector-emitter voltage (see Section 6.2). Realistic values have been obtained as well for the nonlinearity of the Early resistance, which is not modeled in most circuit simulators that are used nowadays.

The assumption that one measured harmonic or intermodulation product is determined by only one coefficient is only valid at low frequencies and when the internal ohmic resistances of the transistor are negligible. When this is not the case, then a harmonic or intermodulation product is determined by more than one coefficient and maybe by more than one basic nonlinearity. In order to obtain values for the nonlinear coefficients in this case, more than one harmonic or intermodulation product must be measured and the measurement results must be combined. However, the measurements reported in this chapter are performed on a large transistor at low frequencies, such that a coefficient can be determined from one single harmonic or intermodulation product.

# 9.2 Principle of the measurements

The idea behind the measurements is to obtain a value for a nonlinearity coefficient with one measured harmonic or intermodulation product, instead of having to combine several measurement results out of which several coefficients can be extracted. If with this idea good measurement results can be obtained, then the measurement setup that is used for the measurements can be used as well in cases where it is not possible to extract one coefficient from every measured harmonic or intermodulation product.

The equations that model the drain current of a MOS transistor or the collector current of a bipolar transistor both express the current as a function of one or more voltages. The derivatives of the current with respect to one voltage can be measured when the appropriate AC voltages are applied. For example, when an AC voltage at low frequencies is applied between the gate and the source of a properly biased MOS transistor while the drain-source voltage and the bulk-source voltage are kept co istant, then the AC drain current is only dependent on the applied AC voltage between gate and source. In other words, if the AC voltage between gate and source is given by

$$v_{in1} = V_{in1}\cos(\omega_1 t) \tag{9.1}$$

and the AC drain current only depends on the gate-source voltage:

$$i_d = g_m \cdot v_{gs} + K_{2g_m} \cdot v_{gs}^2 + K_{3g_m} \cdot v_{gs}^3 + \dots$$
 (9.2)

then the first three harmonics are found to be

fundamental = 
$$g_m V_{in1} \cos(\omega_1 t)$$
 (9.3)

2nd harmonic = 
$$K_{2g_m} \frac{V_{in1}^2}{2} \cos(2\omega_1 t)$$
 (9.4)

$$3 \text{rd harmonic} = K_{3g_m} \frac{V_{in1}^3}{4} \cos(3\omega_1 t)$$
 (9.5)

It is seen that a spectral analysis of the drain current yields the coefficients  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$ . The same principle can be used to determine the coefficients  $g_o$ ,  $K_{2g_o}$ ,  $K_{3g_o}$ ,  $g_{mb}$ ,  $K_{2g_{mb}}$  and  $K_{3g_{mb}}$  for a MOS transistor. For a bipolar transistor, the application of this procedure to the collector current yields the coefficients  $g_m$ ,  $K_{2g_m}$ ,  $K_{3g_m}$ ,  $g_o$ ,  $K_{2g_o}$  and  $K_{3g_o}$ .

More accurate results can be obtained if at the input port under consideration two sine way at different frequencies are applied. Indeed, from Chapter 2 we know that the second-ord intermodulation products are two times higher than the second harmonic of one of the two six waves, while the third-order intermodulation products are three times higher. This is especial useful when the harmonics have amplitudes which come out only slightly above the noise floc

For the determination of coefficients that arise from cross-derivatives, two signals are applications simultaneously at two different input ports. If two excitations

$$v_{in1} = V_{in1}\cos(\omega_1 t)$$
 and  $v_{in2} = V_{in2}\cos(\omega_2 t)$  (9.

are applied, then the harmonics of  $\omega_1$  and  $\omega_2$  will be proportional to nonlinearity coefficients the are proportional to derivatives with respect to one voltage, while the intermodulation production

are proportional to the coefficients that are determined by cross-derivatives. Hereby care must be taken that the two frequencies  $\omega_1$  and  $\omega_2$  are not harmonically related.

As an example, assume that  $v_{in1}$  from equation (9.6) is applied between the base and emitter of a properly biased bipolar transistor, while  $v_{in2}$  is applied between the collector and the emitter. In this case, the amplitude of the second-order intermodulation product at the frequency  $|\omega_1 \pm \omega_2|$ is given by

$$V_{in1}V_{in2}K_{2g_m\&g_o} (9.7)$$

The amplitudes of the third-order intermodulation products are given by

at 
$$|2\omega_1 \pm \omega_2|$$
:  $\frac{3}{4}V_{in1}^2V_{in2}K_{3_{2g_m\&g_o}}$  (9.8)

at 
$$|2\omega_1 \pm \omega_2|$$
:  $\frac{3}{4}V_{in1}^2V_{in2}K_{3_{2g_m\&g_o}}$  (9.8)  
at  $|2\omega_2 \pm \omega_1|$ :  $\frac{3}{4}V_{in1}V_{in2}^2K_{3_{g_m\&2g_o}}$  (9.9)

This principle can be used to determine all coefficients that are determined by derivatives with respect to two voltages. It can be extended to coefficients that are determined by derivatives with respect to three independent voltages such as the coefficient  $K_{3_{g_m\&g_{mb}\&g_o}}$  of a MOS transistor.

#### **Practical applications** 9.3

It is now investigated how the principle explained in the previous section can be applied in practice to determine the nonlinear coefficients of bipolar and MOS transistors by measurements of harmonics and intermodulation products.

A measurement setup for the determination of the nonlinearities of the drain current of a MOS transistor is shown in Figure 9.1. Transistor  $M_1$  is the device under test. Up to three independent AC voltages can be applied:  $v_{in1}$ ,  $v_{in2}$  and  $v_{in3}$ . The AC sources are buffered with operational amplifiers (OA1, OA2 and OA3, respectively). The AC sources are connected in series with DC voltage sources that provide the correct bias for  $M_1$ . When less than three excitations are applied, then the input of every inactive buffer is shorted to AC ground. The drain current is measured by the voltage drop over resistance  $R_L$  by means of an instrumentation amplifier IA1. The shuntshunt feedback configuration with  $R_L$  ensures a low impedance at the drain of transistor  $M_1$ . In this way, the AC voltage at the drain does not change significantly due to the applied signals at the gate or the bulk.

The measurement must be performed at low frequencies, such that no parasitic capacitance plays a role. Also, the loop gain of the opamps in unity-gain configuration must be sufficiently high at the signal frequencies, such that the impedance seen at the output of every buffer is very small.

A similar principle can be used for the measurement of the nonlinearities of a bipolar transistor. The measurement setup is shown in Figure 9.2. The device under test, transistor  $Q_1$ , cannot be biased in the same way as the MOS transistor from Figure 9.1: because of the exponential relationship between the base-emitter voltage and the collector current, a slight change of a voltage applied between base and emitter, for example due to an external disturbance, may cause

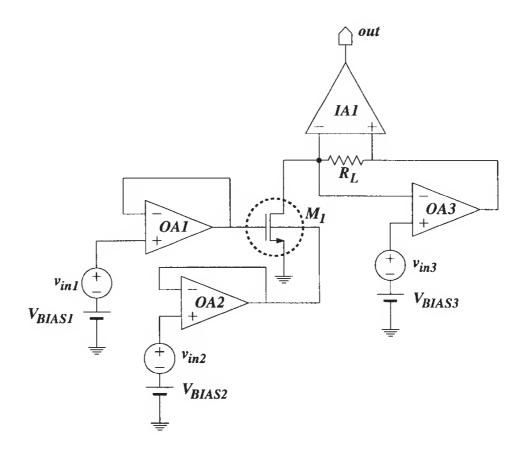


Figure 9.1: Measurement setup for the determination of the nonlinearities of the drain current of a MOS transistor.

a significant change of the collector current. Instead, a base current is inserted into the bipolar transistor under test using an external DC current source. The capacitor  $C_1$  decouples the DC voltage at the base of  $Q_1$  from the AC excitation  $v_{in1}$ .

For both a MOS and a bipolar transistor the above measurement setup can be used to determine immediately the nonlinear coefficients as explained in Section 9.2, if the internal ohmic resistances of the transistor can be neglected. However, if, for example, measurements an performed on a bipolar transistor with a significant base resistance, then this base resistance influences the measured harmonics<sup>1</sup>. In the presence of ohmic resistances the formulas (9.2) through (9.9) are not correct anymore. Instead, the different measured harmonics and intermodulation products can depend on several nonlinearity coefficients as well as on several small-sign parameters. In that case, the different nonlinearity coefficients should be determined by combining measured values of different harmonics or intermodulation products. These values can dentified with symbolic expressions for the measured harmonics and intermodulation products. These expressions can be derived with the method described in Chapter 5. In this way, of obtains a set of equations with several unknown nonlinearity coefficients. The solution of the set then gives the values of the required nonlinearity coefficients. In Section 9.4 measurement

<sup>&</sup>lt;sup>1</sup>This is also the case if a series resistance is inserted between the DC current source and the base of the transit to measure the nonlinearity of the base current.

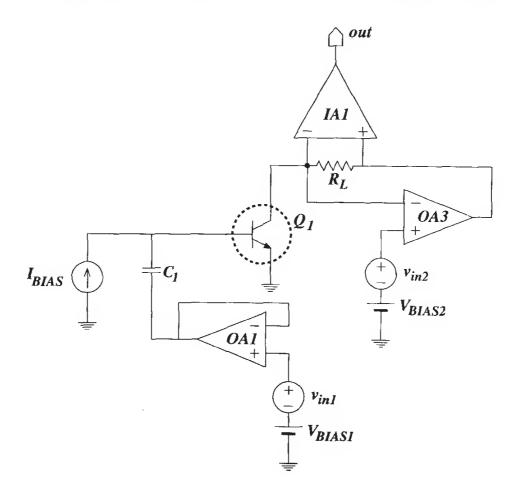


Figure 9.2: Measurement setup for the determination of the nonlinearities of the collector current of a bipolar transistor.

are presented on large transistors in which the ohmic resistances are negligible, such that the principle of Section 9.2 can be used.

A final remark must be made about the temperature of the device under test. When nonlinear coefficients are measured over a wide range of the bias current, varying from very low to very high, then precautions must be taken to keep the temperature in the device constant. Else, the temperature dependence of the different coefficients is measured together with the coefficients themselves. The measurements that are reported in Section 9.4 are performed over a small range of bias currents such that the temperature is nearly constant. This will be illustrated with the measurement results in the next section.

# 9.4 Measurement results

Measurements of harmonics and intermodulation products have been performed on a large bipolar transistor with an emitter area of  $2\mu m$  by  $500\mu m$ . Due to this large emitter area the ohmic base, emitter and collector resistances can be neglected.

The measurement setup is the one shown in Figure 9.2. The load resistor  $R_L$  is a metal film resistor. Its value depends on the signal levels that are measured. It is bridged by a small capacitor in order to compensate for a pole at the negative input of the buffer. For the operational amplifiers the OP27 [AD 88] is used. The instrumentation amplifier is the AD 524 from Analog Devices [AD 88]. The output of the instrumentation amplifier is fed into the dynamic signal analyzer HP 3562A which measures the harmonics and the intermodulation products. This signal analyzer also has a voltage source with a total harmonic distortion of -60 dB at 1V below 10 kHz. This source is used as one of the two AC sources in the measurement setup. The other AC source comes from the distortion analyzer HP 339A which has a total harmonic distortion of -100 dBat 1V below 10kHz. The amplitudes of the sine waves are determined by the requirement that the harmonics and intermodulation products of interest should have an amplitude that is higher than the noise floor. The measurement setup has been checked first with metal film resistors instead of a transistor. Then the highest signal levels that are used in the measurements have been applied. With these conditions no harmonics and intermodulation products have been detected with the dynamic signal analyzer. For each signal level that is used, it has been checked whether the amplitude of a harmonic or intermodulation product changes in the same way as predicted in Section 9.2. This means that if the signal level at one of the two input ports is increased with 1 dB, then the harmonic or intermodulation product that is proportional to the nth power (n = 1, 2 or 3) of the input amplitude, should increase with  $n \, dB$ . If this were not the case then this means that a harmonic or intermodulation product of order n is also determined by nonlinear behavior of order higher than n and the applied signal level must be lowered. The frequencies that are used for the two excitations are 985Hz and 875Hz.

The bipolar transistor has been biased in its forward active region far from the high-injection region: the forward knee current [Lak 94] is about 0.1A. Also, temperature changes only slightly in this region, as will be confirmed by measurements.

With the above measurement conditions the different derivatives of the collector current can be computed from the measured harmonics and intermodulation products. The results are presented in the next sections.

# 9.4.1 Derivatives of $i_C$ with respect to $v_{BE}$

Far from the high-injection region and at a constant temperature, the simple exponential law for the collector current as a function of the base-emitter voltage is valid. In this case, it has been shown in Section 3.2 that

$$\frac{I_C}{g_m} = \frac{g_m}{2K_{2g_m}} = \frac{K_{2g_m}}{3K_{3g_m}} = V_t \tag{9.1}$$

Since  $V_t = kT/q$ , the measurement of  $i_C$ ,  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$  can give an idea of the average temperature in the device. The value of  $I_C$  has been obtained with DC measurements, the value for  $g_m$ ,  $K_{2g_m}$  and  $K_{3g_m}$  have been extracted from harmonics in the collector current. To the purpose, a sinusoidal excitation  $v_{in1}$  (see Figure 9.2) is applied with a frequency of 985Hz an amplitude of 7mV. The source  $v_{in2}$  is not used in this measurement. The fundament

component of the collector current is proportional to the transconductance  $g_m$ , while the second and the third harmonic are proportional to  $K_{2g_m}$  and  $K_{3g_m}$ , respectively.

The ratios of equation (9.10) as they are deduced from the measurement results are shown in Figure 9.3. In the range of the  $v_{BE}$  values depicted in Figure 9.3, the collector current changes

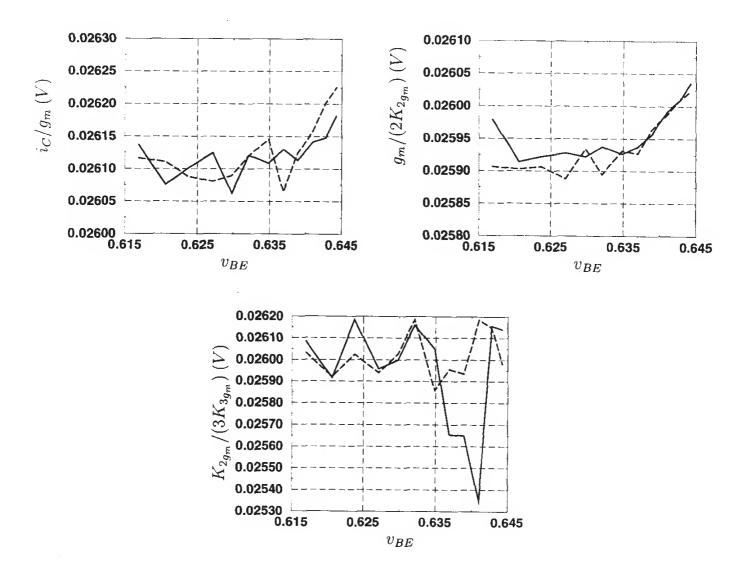


Figure 9.3: Ratios of subsequent derivatives of  $i_C$  with respect to  $v_{BE}$  as extracted from measurements. The solid lines correspond to  $v_{CE}=2.5V$ , the dashed lines to  $v_{CE}=4.5V$ .

from  $20\mu A$  to  $25\mu A$ . It is seen that the different ratios match very well to the predicted value of equation (9.10). Also, it can be concluded that the temperature does not change significantly over the range of values for  $v_{BE}$  and  $v_{CE}$ . This also holds for a wider range of  $v_{BE}$  values: measurements from  $2\mu A$  to  $50\mu A$  yield results with a comparable accuracy. Hence, it can be concluded that the results indicate that with the measurement technique harmonics up to order three can be measured accurately. Next, some interesting results from measurements of intermodulation products are discussed.

### 9.4.2 The nonlinearity of the Early resistance

In circuit simulators a bipolar transistor is most often modeled with a constant Early voltage [Lak 94, Hspi 96]. This means that derivatives of the collector current obtained by differentiating more than once with respect to the collector-emitter voltage, are zero. Consequently, coefficients such as  $K_{2g_o}$ ,  $K_{3g_o}$  and  $K_{3g_m\&2g_o}$  are zero. However, this is not correct. In many applications this error does not cause significant deviations between simulations and measurements. However, in [Opt 89] it is shown that for ultra-low distortion applications the performance of some circuits is limited by the nonlinearity of the CMRR, which is partially determined by the nonlinearity of the Early resistance.

For the measurement of the coefficients  $K_{2g_o}$  and  $K_{3g_o}$  the voltage source  $v_{in1}$  in Figure 9.2 has been removed and put in series with  $v_{in2}$ . The amplitude of both sources has been made identical. The second- and third-order intermodulation products yield values for  $K_{2g_o}$  and  $K_{3g_o}$ , respectively. Instead of computing  $K_{2g_o}$  and  $K_{3g_o}$  from intermodulation products, they could have been obtained as well from harmonics, in a similar way as  $K_{2g_m}$  and  $K_{3g_m}$  have been obtained in the previous section. However, in this case the amplitudes of the harmonics are quite close to the noise floor, such that the accuracy on the values obtained from intermodulation products instead of harmonics is significantly higher.

The set of values for  $K_{3g_o}$  that have been measured over the range of  $v_{CE}$  values, have been integrated numerically over this range. Since

$$K_{2g_o} = \frac{1}{2} \frac{\partial^2 i_C}{\partial v_{CE}^2} \tag{9.11}$$

and

$$K_{3g_o} = \frac{1}{6} \frac{\partial^3 i_C}{\partial v_{CE}^3},\tag{9.12}$$

it is clear that

$$K_{2g_o}(v_{CE2}) = 3 \int_{v_{CE1}}^{v_{CE2}} K_{3g_o} dv_{CE} - K_{2g_o}(v_{CE1})$$
(9.13)

In this equation,  $K_{2g_o}$  is evaluated at a value  $v_{CE2}$ , which is the collector-emitter voltage at a value inside the range of  $v_{CE}$  that is used in the measurements. The value  $K_{2g_o}(v_{CE1})$  is the measured value at the smallest of the values for  $v_{CE}$  (in this case 2.5V). Hence, a numerical integration of the measurement results for  $K_{3g_o}$  can yield a value for  $K_{2g_o}$ . This integral together with the measured value of  $K_{2g_o}$  is shown in Figure 9.4 for a  $v_{BE}$  value of 0.637V, that corresponds to a collector current of about  $50\mu A$ . It is seen that a good agreement is obtained between the two values. This is an extra confirmation for the quality of the measurements.

Figure 9.5 shows the measured value for  $g_o$  compared to the value obtained by the numerical integration of the measured data for  $K_{2g_o}$ , as well as the measured value for  $i_C$  compared to the value obtained by the numerical integration of the measured data for  $g_o$ . Both plots have been produced for a  $v_{BE}$  value of 0.637V, that corresponds to a collector current of about  $50\mu A$ . The

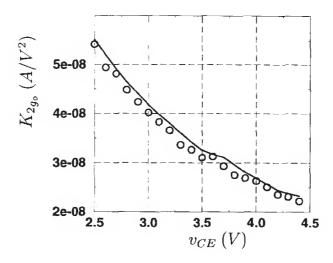


Figure 9.4: Measured value for  $K_{2g_o}$  (circles) and the value of  $K_{2g_o}$  obtained by integration from  $K_{3g_o}$  (solid line).

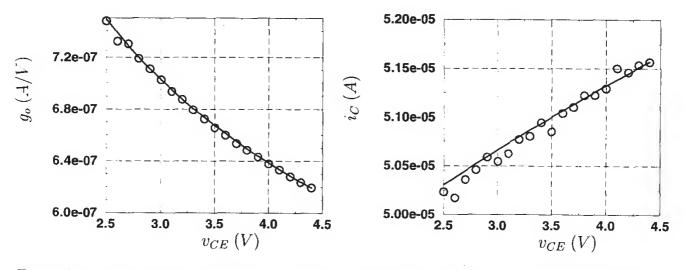


Figure 9.5: Left: measured value for  $g_o$  (circles) compared to the value obtained by the numerical integration of the measured data for  $K_{2g_o}$  (solid line). Right: measured value for  $i_C$  (circles) compared to the value obtained by the numerical integration of the measured data for  $g_o$  (solid line).

collector current has been measured at DC. It is seen that in both cases a good agreement is obtained.

It must be remarked that at low frequencies there is more than one intermodulation product from which  $K_{2g_o}$  and  $K_{3g_o}$  can be extracted. For example,  $K_{2g_o}$  can be extracted from the second-order intermodulation product at the frequency which is the sum of the two applied frequencies, as well as at the difference frequency. This should give the same results. In practice, deviations of 1.5% have been measured.

Finally, it should be noticed that the measurement of higher-order derivatives, followed by an

integration to obtain the lower-order derivatives, yields more accurate results than the measurement or modeling of lower-order derivatives followed by a numerical differentiation in order to obtain higher-order derivatives.

### 9.4.3 Cross-derivatives of $i_C$

Finally, measurement results are presented for the cross-derivatives of  $i_C$ , which are derivatives of  $i_C$  with respect to both  $v_{BE}$  and  $v_{CE}$ . These figures are obtained by measuring intermodulation products that are caused by sources applied at two different input ports, as depicted in Figure 9.2. The excitation  $v_{in1}$  is a sine wave at  $\omega_1 = 2\pi \times 985Hz$  with an amplitude of 7~mV, while  $v_{in2}$  is a sine wave at  $\omega_2 = 2\pi \times 875Hz$  with an amplitude of 300~mV. In this way, the following data can be obtained:

response at 
$$|\omega_1 \pm \omega_2|$$
 yields  $K_{2_{q_m} \& q_q}$  (9.14)

response at 
$$|2\omega_1 \pm \omega_2|$$
 yields  $K_{3_{2q_m \& q_0}}$  (9.15)

response at 
$$|\omega_1 \pm 2\omega_2|$$
 yields  $K_{3q_m\&2q_o}$  (9.16)

Of course, from the harmonics of the collector current the coefficients  $g_m$ ,  $K_{2g_m}$ ,  $K_{3g_m}$ ,  $g_o$ ,  $K_{2g_o}$  and  $K_{3g_o}$  can be extracted. However, it is more accurate to measure these coefficients by applying two sources in series at the same input port, as performed in Section 9.4.2.

Just as in the previous section, the quality of the measurements can be controlled by comparing a measured lower-order value to the integral of a higher-order derivative. For example, an integration of  $K_{2g_m\&g_o}$  over the range of  $v_{CE}$  values yields  $g_m$  as a function of  $v_{CE}$ , which can be compared to measured values of  $g_m$ . This comparison is shown in Figure 9.6. Again a good agreement is found between the directly measured data and the integrated data.

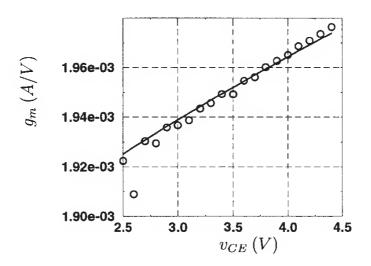


Figure 9.6: Measured value for  $g_m$  (circles) and the value of  $g_m$  obtained by integration from  $K_{2g_m\&g_o}$  (solid line).

# 9.5 Summary

In this chapter a principle has been presented for the measurement of the first-, second- and thirdorder coefficients that characterize basic nonlinearities of the drain current or collector current of
a MOS or bipolar transistor. The idea is to measure harmonics or intermodulation products that
only depend on one or just a few coefficients. This is a much more accurate approach than the
measurement or modeling of the current after which the derivatives of different order are obtained
by a differentiation of the model equation or of the measured data for the current. The principle
has been tested with measurements of the different coefficients that characterize the nonlinearity
of the collector current of a large bipolar transistor. The quality of the measurements is good:
this is shown by the close agreement between measured data of lower-order coefficients and the
numerical integration of higher-order coefficients. Measurement results have been obtained for
the nonlinearity of the Early resistance, which have not been reported previously.

The measured data for the second- and third-order coefficients can be used in parameter extraction or device modeling for an accurate modeling of the higher-order nonlinearities of a transistor.

# **Bibliography**

The following abbreviations have been used:

**CICC** 

DAC

**ECCTD** 

**ICCAD** 

**IEDM** 

**ISCAS** 

**ISSCC** 

Custom Integrated Circuits Conference

Design Automation Conference

European Conference on Circuit Theory and

Design

International Conference on Computer-Aided

Design

International Electron Devices Meeting

International Symposium on Circuits and

Systems

International Solid-State Circuits Conference

- [Abid 95] A. Abidi, "Direct-conversion radio transceivers for digital communications," *IEEE J. Solid-State Circuits*, Vol. 30, No. 12, pp. 1399-1410, December 1995.
- [Abra 76] H. Abraham and R. Meyer, "Transistor design for low distortion at high frequencies," *IEEE Trans. Electron Devices*, Vol. ED-23, No. 12, pp. 1290–1297, December 1976.
- [Ald 73] G.E. Alderson and P.M. Lin, "Computer generation of symbolic network functions A new theory and implementation," *IEEE Trans. Circuit Theory*, Vol. CT-20, No. 1, pp. 48–56, 1973.
- [AD 88] Linear products databook, Analog Devices 1988.
- [Anto 88] P. Antognetti, G. Massobrio et al., Semiconductor device modeling with SPICE. McGraw-Hill, 1988.
- [Apr 72] T. Aprille and T. Trick, "Steady-state analysis of nonlinear circuits with periodic inputs," *Proc. IEEE*, Vol. 60, No. 1, pp. 108–114, January 1972.
- [Aro 89] N. Arora and L. Richardson, "MOSFET modeling for circuit simulation," chapter 6 in "VLSI electronics microstructure science, Vol. 18: Advanced MOS Device Physics, edited by N. Einspruch and G. Gildenblat, Academic Press, 1989.
- [Ash 96] K. Ashby, I. Koullias, W. Finley, J. Bastek and S. Moinan, "High Q inductors for wireless applications in a complementary silicon bipolar process," *IEEE J. Solid-State Circuits*, Vol. 31, No. 31, pp. 4–9, January 1996.
- [Asti 96] S. Asti, T. Cavioni, A. Neviani, P. Pavan, M. Stival, L. Vendrame and E. Zanoni, "Analysis of charge storage in the base of bipolar transistors and its influence on the parasitic resistance adopting an eight terminal Kelvin test structure," *Proc. IEEE International Conference on Microelectronic Test Structures 1996*, pp. 91–96.
- [Ath 75] D. Atherton, Nonlinear control engineering describing function analysis. Van Nostrand-Reinhold, New York 1975.
- [Bas 95] J. Bastos et al., "Mismatch characterization of small size MOS transistors," Proc. IEEE International Conference on Microelectronic Test Structures 1995, pp. 271–276.
- [Baum 70] G. Baum and H. Beneking, "Drift velocity saturation in MOS transistors," *IEEE Trans. Electron Devices*, Vol. ED-17, pp. 481–482, June 1970.

- [Bedr 71] E. Bedrosian and S. O. Rice, "The output properties of Volterra systems (nonlinear systems with memory) driven by harmonic and Gaussian inputs," *Proc. IEEE*, Vol. 59, No. 12, pp. 1688–1707, December 1971.
- [Borre 97] M. Borremans and M. Steyaert, "A 2V, low power, single-ended 1 *GHz* CMOS direct upconversion mixer," *Proc. CICC*, 1997, pp. 517–520.
- [Boyd 84] S. Boyd, L. O. Chua, and C. A. Desoer, "Analytic foundations of Volterra series," *IMA J. Math. Control and Information*, Oxford University Press, pp. 243–282, 1984.
- [BSIM 95] BSIM3v3 User's Manual, Dept. of Electrical Engineering and Computer Science, University of Berkeley, 1995.
- [Buss 74] J. Bussgang, L. Ehrman, and J. Graham, "Analysis of nonlinear systems with multiple inputs," *Proc. IEEE*, Vol. 62, No. 8, pp. 1088–1118, August 1974.
- [Chan 84] T.Y. Chan, P.K. Ko and C. Hu, "A simple method to characterize substrate current in MOSFET's," *IEEE Electron Device Letters*, Vol. EDL-4, No. 12, pp. 505–507, December 1984.
- [Chang 97] Z.Y. Chang et al., "A highly linear CMOS G<sub>m</sub>-C bandpass filter with on-chip frequency tuning," *IEEE J. Solid-State Circuits*, Vol. 32, No. 3, pp. 388–397, March 1997.
- [Chen 65] W.K. Chen, "Topological analysis for active networks," *IEEE Trans. Circuit Theory*, Vol. CT-12, No. 1, pp. 85–91, March 1965.
- [Cheng 96] Y. Cheng, M.-C. Jeng, Z. Liu, K. Chen, M. Chan, C. Hu and P.-K. Ko, "An investigation on the robustness, accuracy and simulation performance of a physics-base deep-submicronmeter BSIM model for analog/digital circuit simulation," *Proc. CICC* 1996, pp. 15.1.1–15.1.4.
- [Chis 73] S. Chisholm and L. Nagel, "Efficient computer simulation of distortion in electronic circuits," *IEEE Trans. Circuit Theory*, Vol. CT-20, No. 6, pp. 742–745, November 1973.
- [Chiu 68] T. L. Chiu and C. T. Sah, "Correlation experiments with a two-section model theory of the saturation drain conductance of MOS transistors," *Solid-State Electronics*, Vol. 11, pp. 1149–1163, 1968.
- [Chiu 92] T.-Y. Chiu, P. K. Tien, J. Sung and T.-Y. Mark Liu, "A new analytical model and the impact of base charge storage on base potential distribution, emitter current crowding and base resistance," *Proc. IEDM*, 1992, pp. 573–576.
- [Chow 92a] H.-C. Chow and W.-S. Feng, "An improved analytical model for short-channed MOSFET's," *IEEE Trans. Electron Devices*, Vol. 39, No. 11, pp. 2626–2629, November 1992.

- [Chow 92b] H.-C. Chow, W.-S. Feng and J. Kuo, "An improved analytical short-channel MOS-FET model valid in all regions of operation for analog/digital circuit simulation," *IEEE Trans. Computer-Aided Design*, Vol. 11, No. 12, pp. 1522–1528, December 1992.
- [Chua 75] L.O. Chua and P.-M. Lin, Computer-aided analysis of electronic circuits: algorithms & computational techniques. Prentice-Hall, 1975.
- [Chua 79a] L.O. Chua and C.-Y. Ng, "Frequency-domain analysis of nonlinear systems: general theory," *IEE J. Electronic Circuits and Systems*, Vol. 3, No. 4, pp. 165–185, July 1979.
- [Chua 79b] L.O. Chua and C.-Y. Ng, "Frequency-domain analysis of nonlinear systems: formulation of transfer functions," *IEE J. Electronic Circuits and Systems*, Vol. 3, No. 6, pp. 257–269, November 1979.
- [Chua 82] L.O. Chua and Y.-S. Tang, "Nonlinear oscillation via Volterra series," *IEEE Trans. Circuits and Systems*, Vol. CAS-29, No. 3, pp. 150–168, March 1982.
- [Chua 87] L.O. Chua, C. Desoer and E. Kuh, *Linear and nonlinear circuits*. Mc Graw-Hill, 1987.
- [Coat 58] C.L. Coates, "Flow-graph solutions of linear algebraic equations," *IRE Trans. Circuit Theory*, Vol. 6, pp. 170–187, June 1959.
- [Coen 80] R. Coen and R. Muller, "Velocity of surface carriers in inversion layers on silicon," *Solid-State Electronics*, Vol. 23, pp. 35–40, 1980.
- [Coop 81] J. Cooper and D. Nelson, "Measurement of the high-field drift velocity of electrons in inversion layers on silicon," *IEEE Electron Device Letters*, Vol. EDL-2, No. 7, pp. 171–173, July 1981.
- [Crols 95b] J. Crols and M. Steyaert, "A single-chip 900 MHz CMOS receiver front-end with a high performance low-IF topology," *IEEE J. Solid-State Circuits*, Vol. 30, No. 12, pp. 1483–1492, December 1995.
- [Crols 95a] J. Crols and M. Steyaert, "A 1.5 *GHz* highly linear CMOS downconversion mixer," *IEEE J. Solid-State Circuits*, Vol. 30, No. 7, pp. 736–742, July 1995.
- [Dam 93] D. D'Amore and W. Fornaciari, "A SPICE-based approach to steady-state circuit analysis," *Int. J. Circuit Theory and Applications*, Vol. 21, pp. 437–442, 1993.
- [Daw 97] G. Dawe, J.-M. Mourant and A.P. Brokaw, "A 2.7V DECT RF transceiver with integrated VCO," *Proc. ISSCC 1997*, pp. 308–309.
- [DeVr 96] L. De Vreede, H. de Graaff, K. Mouthaan, M. de Kok, J. Tauritz and R. Baets, "Advanced modeling of distortion effects in bipolar transistors using the Mextram model," *J. Solid-State Circuits*, Vol. 31, No. 1, pp. 114–121, January 1996.

- [Dun 91] L. Dunlop, "An efficient MOSFET current model for analog circuit simulation subthreshold to strong inversion," *J. Solid-State Circuits*, Vol. 25, No. 2, pp. 616–619, April 1990.
- [Eldo 91] ELDO model equations, Anacad Computer Systems GmbH., 1991.
- [El-Man 77] Y. El-Mansy and A. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Devices*, Vol. ED-24, pp. 71–83, 1976.
- [Enz 95] C. Enz, F. Krummenacher and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Kluwer J. Analog Integrated Circuits and Signal Processing*, Vol. 8, pp. 83–114, 1995.
- [Feld 96] P. Feldmann, B. Melville and D. Long, "Efficient frequency domain analysis of large nonlinear analog circuits," *Proc. CICC*, 1996, pp. 461–464.
- [Fern 91b] F.V. Fernández, A. Rodríguez-Vázquez and J.L. Huertas, "Interactive ac modeling and characterization of analog circuits via symbolic analysis," *Kluwer J. Analog Integrated Circuits and Signal Processing*, Vol. 1, pp. 183–208, November 1991.
- [Fern 93a] F.V. Fernández, A. Rodríguez-Vázquez and J.L. Huertas, "Formula approximation for flat and hierarchical symbolic analysis," *Kluwer J. Analog Integrated Circuits and Signal Processing*, Vol. 3, No. 1, pp. 43–58, January 1993.
- [Foty 96] D. Foty, MOSFET modeling with SPICE: principles and practice. Prentice-Hall, 1996.
- [Froh 68] D. Frohman-Bentchkowsky, "On the effect of mobility variation on MOS device characteristics," *IEEE Proceedings*, pp. 217–218, February 1968.
- [Froh 69] D. Frohman-Bentchkowsky and A. Grove, "Conductance of MOS transistors in saturation," *IEEE Trans. Electron Devices*, Vol. ED-16, No. 1, pp. 108–113, January 1969.
- [Fuse 92] T. Fuse *et al.*, "A physically based base pushout model for submicrometer BJT's in the presence of velocity overshoot," *IEEE Trans. Electron Devices*, Vol. 39, No. 2, pp. 396–403, February 1992.
- [Fuse 95] T. Fuse and Y. Sasaki, "An analysis of small-signal and large signal base resistances for submicrometer BJT's," *IEEE Trans. Electron Devices*, Vol. 42, No. 3, pp. 534, 539, March 1995.
- [Gar 87a] S. Garverick and C. Sodini, "A simple model for scaled MOS transistors that include field-dependent mobility," *IEEE J. Solid-State Circuits*, Vol. SC-22, No. 1, pp. 111-114, February 1987.

- [Gar 87b] S. Garverick and C. Sodini, "Large-signal linearity of scaled MOS transistors," *IEEE*J. Solid-State Circuits, Vol. SC-22, No. 2, pp. 282–286, April 1987.
- [Getr 76] I. Getreu, Modeling the bipolar transistor. Tektronix, Inc., 1976.
- [Giel 89] G. Gielen, H. Walscharts, and W. Sansen, "ISAAC: a symbolic simulator for analog integrated circuits," *IEEE J. Solid-State Circuits*, Vol. 24, No. 6, pp. 1587–1597, December 1989.
- [Giel 91] G. Gielen and W. Sansen, Symbolic analysis for automated design of analog integrated circuits. Kluwer Academic Publishers, 1991.
- [Giel 94a] G. Gielen, P. Wambacq and W. Sansen, "Symbolic analysis methods and applications for analog circuits: a tutorial overview," *Proc. IEEE*, Vol. 82, No. 2, pp. 287–304, February 1994.
- [Gilb 68] B. Gilbert, "A new wide-band amplifier technique," *IEEE J. Solid-State Circuits*, Vol. SC-3, No. 4, pp. 353–365, December 1968.
- [Gilb 74] B. Gilbert, "A high-performance monolithic multiplier using active feedback," *IEEE J. Solid-State Circuits*, Vol. SC-9, No. 6, pp. 364–373, December 1974.
- [Gilb 96] B. Gilbert, "Design considerations for BJT active mixers," chapter 23 in "Low-power HF Microelectronics: a unified approach", edited by G. Machado, IEE 1996.
- [Gopi 90] V. Gopinathan, Y. Tsividis, K.-S. Tan, and R. Hester, "Design considerations for high-frequency continuous-time filters and implementation of an antialiasing filter for digital video" *IEEE J. Solid-State Circuits*, Vol. 25, No. 6, pp. 1368–1378, December 1990.
- [Gow 91] S. Gowda, B. Sheu and J. Cable, "An accurate MOS transistor model for submicron VLSI circuits BSIM\_plus," *Proc. CICC 1991*, pp. 23.2.1–23.2.4.
- [Gow 93] S. Gowda, B. Sheu and C.-H. Chang, "Advanced VLSI circuit simulation using the BSIM\_plus model," *Proc. CICC 1993*, pp. 14.3.1–14.3.5.
- [Gray 93] P. Gray and R. Meyer, Analysis and design of analog integrated circuits, 3rd edition. New York: J.Wiley, 1993.
- [Groen 94] G. Groenewold and J. Lubbers, "Systematic distortion analysis for MOSFET integrators with use of a new MOSFET model," *IEEE Trans. Circuits and Systems II:*Analog and Digital Signal Processing, Vol. 41, No. 9, pp. 569–580, September 1994.
- [Grot 84] T. Grotjohn and B. Hoefflinger, "A parametric short-channel MOS transistor model for subthreshold and strong inversion current," *IEEE Trans. Electron Devices*, Vol. ED-31, No. 2, pp. 234–246, February 1984.

- [Gueb 83] P. Guebels and F. Van De Wiele, "A small geometry MOSFET model for CAD applications," *Solid-State Electronics*, Vol. 26, pp. 267–273, April 1983.
- [Hamel 96] J. Hamel, "Compact modeling of the influence of emitter stored charge on the high frequency small signal AC response of bipolar transistors using quasi-static parameters," *IEEE J. Solid-State Circuits*, Vol. 31, No. 1, pp. 106–113, January 1996.
- [Har 96] Application Note 9618.2: Using the Prism HFA3624 evaluation board, Harris Semiconductor, Florida, 1996.
- [Hass 89] M. Hassoun and P.M. Lin, "A new network approach to symbolic simulation of large-scale networks", *Proc. ISCAS*, 1989, pp. 806–809.
- [Herm 91] R. Herman, A. Chao, c. Mason and J. Pulver, "An integrated GPS receiver with synthesizer and downconversion functions," *IEEE MTT-S Digest*, 1991, pp. 883–886.
- [Hspi 96] HSPICE User's Manual Release 96.1, Meta-Software Inc., 1996.
- [Hsu 83] F.-C. Hsu, R. Muller and C. Hu, "A simplified model of short-channel MOSFET characteristics in the breakdown mode," *IEEE Trans. Electron Devices*, Vol. ED-30, pp. 571–576, 1983.
- [Hsu 89] F.-C. Hsu, "Hot-carrier-resistant structures," chapter 4 in "VLSI electronics microstructure science, Vol. 18: Advanced MOS Device Physics", edited by N. Einspruch and G. Gildenblat, Academic Press, 1989.
- [Hsu 94] J.J. Hsu and C. Sechen, "DC small signal symbolic analysis of large analog integrated circuits," *IEEE Trans. Circuits and Systems I: fundamental theory and applications*, Vol. 41, No. 12, pp. 817–828, December 1994.
- [Hua 87] G.-S. Huang and C.-Y. Wu, "An analytic I-V model for lightly doped drain (LDD) MOSFET devices," *IEEE Trans. Electron Devices*, Vol. EDL-34, No. 6, pp. 1311-1322, June 1987.
- [Hua 90] C.-L. Huang and G. Gildenblat, "Measurements and modeling of the n-channel MOS FET inversion layer mobility and device characteristics in the temperature range 60 300 K," *IEEE Trans. Electron Devices*, Vol. 37, No. 5, pp. 1289–1300, May 1990.
- [Hua 92] J. Huang, Z. Liu, M. Jeng, P. Ko and C. Hu, "A physical model for MOSFET output resistance," *Proc. IEDM*, 1992, pp. 569–572.
- [Hua 93] J. Huang, Z. Liu, M. Jeng, P. Ko and C. Hu, "A robust physical and predictive mode for deep-submicrometer MOS circuit simulation," *Proc. CICC*, 1993, pp. 14.2.1, 14.2.3.
- [Hue 89] L. Huelsman, "Personal computer symbolic analysis programs for undergraduate gineering course," *Proc. ISCAS*, 1989, pp. 798–801.

- [Hull 96] C. Hull, J. Tham and R. Chu, "A direct-conversion receiver for 900 MHz (ISM band) spread-spectrum digital cordless telephone," *IEEE J. Solid-State Circuits*, Vol. 31, No. 12, pp. 1955–1963, December 1996.
- [Iwai 85] H. Iwai, M. Pinto, C. Rafferty, J. Oristian and R. Dutton, "Velocity saturation effect on short-channel MOS transistor capacitance," IEEE Electron Device Letters, Vol. EDL-6, pp. 120–122, March 1985.
- [Iked 72] H. Ikeda, "An elegant method for measuring MOST drain-source conductance in the saturated current region," *IEEE Trans. Instrumentation and Measurement*, Vol. IM-21, pp. 234–236, August 1972.
- [Joard 95] K. Joardar, "A new approach for direct observation of base width modulation in vertical bipolar transistors," *IEEE Trans. Electron Devices*, Vol. 42, No. 12, pp. 2189–2196, December 1995.
- [Kara 96] A. Karanicolas, "A 2.7-V 900 MHz CMOS LNA and mixer," IEEE J. Solid-State Circuits, Vol. 31, No. 12, pp. 1939–1944, December 1996.
- [Kend 86] J. Kendall and A. Boothroyd, "A two-dimensional analytical threshold voltage model for MOSFET's with arbitrarily doped substrate," *IEEE Electron Device Letters*, Vol. EDL-7, p. 407, 1986.
- [Khad 74] A. Khadr and R. Johnston, "Distortion in high-frequency FET amplifiers," *IEEE J. Solid-State Circuits*, Vol. SC-9, No. 4, pp. 180–189, August 1974.
- [King 97] P. Kinget and M. Steyaert, "A 1-GHz CMOS up-conversion mixer," *IEEE J. Solid-State Circuits*, Vol. 32, No. 3, pp. 370–376, March 1997.
- [Klaas 76] F. Klaassen, "A MOS model for computer-aided design," *Philips Research Report*, Vol. 31, pp. 71–83, 1976.
- [Klaas 96] D. Klaassen, "Compact modelling of submicron CMOS," *Proc. ESSCIRC*, 1996, pp. 41–46.
- [Khour 87] J. Khoury and Y. Tsividis, "Analysis and compensation of high-frequency effects in integrated MOSFET-C continuous-time filters", *IEEE Trans. Circuits and Systems*, Vol. 36, No. 8, pp. 862–875, August 1987.
- [Koh 89] P. K. Ko, "Approaches to scaling," chapter 1 in "VLSI electronics microstructure science, Vol. 18: Advanced MOS Device Physics", edited by N. Einspruch and G. Gildenblat, Academic Press, 1989.
- [Kon 88] A. Konczykowska and M. Bon, "Automated design software for switched-capacitor ICs with symbolic simulator SCYMBAL," *Proc. ACM/IEEE DAC*, 1988, pp. 363–368.

- [Kund 90] K. Kundert, J. White, and A. Sangiovanni-Vincentelli, Steady-state methods for simulating analog and microwave circuits. Kluwer Academic Publishers, 1990.
- [Kuo 73] Y. Kuo, "Distortion analysis of bipolar transistor circuits," *IEEE Trans. Circuit Theory*, Vol. CT-20, No. 6, pp. 709–716, November 1973.
- [Kuo 77] Y. Kuo, "Frequency-domain analysis of weakly nonlinear networks, "canned" Volterra Analysis, part 1," *IEEE Circuits and Systems*, Vol. 11, No. 4, pp. 2–8, August 1977.
- [Lak 94] K. Laker and W. Sansen, Design of analog integrated circuits and systems. Mc Graw-Hill, 1994.
- [Lar 93] T. Larsen, "Determination of Volterra transfer functions of non-linear multiport networks," J. Wiley Int. J. Circuit Theory and Applications, Vol. 21, pp. 107–131, 1993.
- [Lee 73] H. Lee, "An analysis of the threshold voltage for short-channel IGFET's," *Solid-State Electronics*, Vol. 16, p. 1407, 1973.
- [Lew 89] A. Lewis and J. Chen, "Current trends in MOS process integration," chapter 2 in "VLSI electronics microstructure science, Vol. 18: Advanced MOS Device Physics, edited by N. Einspruch and G. Gildenblat, Academic Press, 1989.
- [Liang 86] M.-S. Liang, J. Y. Choi, P.-K. Ko and C. Hu, "Inversion-layer capacitance and mobility of very thin gate-oxide MOSFET's," *IEEE Trans. Electron Devices*, Vol. ED-33, No. 3, pp. 409–413, March 1986.
- [Liu 82] S. Liu and L. Nagel, "Small-signal MOSFET models for analog circuit design," *IEEE J. Solid-State Circuits*, Vol. SC-17, No. 6, pp. 983–998, December 1982.
- [Liu 93] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko and Y.C. Cheng, "Threshold voltage model for deep-submicrometer MOSFET's," *IEEE Trans. Electron Devices*, Vol. 40, No. 1, pp. 86–95, January 1993.
- [Long 95] J. Long and M. Copeland, "A 1.9 *GHz* low-voltage silicon bipolar receiver front-end for wireless personal communications systems," *IEEE J. Solid-State Circuits*, Vol. 30 No. 12, pp. 1438–1448, December 1995.
- [Long 95] J. Long and M. Copeland, "The modeling, characterization and design of monolithic inductors for silicon RF IC's," *IEEE J. Solid-State Circuits*, Vol. 32, No. 3, pp. 357-369, March 1997.
- [Lot 68] H. Lotsch, "Theory of nonlinear distortion produced in a semiconductor diode," *IEEE Trans. Electron. Devices*, Vol. ED-15, pp. 294–307, May 1968.
- [Maas 88] S. Maas, Nonlinear microwave circuits. Artech House, 1988.

- [Mc And 95] C. McAndrew et al., "VBIC95, the vertical bipolar inter-company model," *IEEE J. Solid-State Circuits*, Vol. 31, No. 10, pp. 1476–1483, October 1996.
- [Macs 87] MACSYMA User's Guide, Symbolics Inc., June 1987.
- [Madi 96] M. Madihian, K. Imai, H. Yoshida, Y. Kinoshita and T. Yamazaki, "L-C-band low-voltage BiCMOS MMIC's for dual-mode cellular-LAN applications," *IEEE Trans. Microwave Theory and Techn.*, Vol. 44, No. 11, pp. 2025–2030, November 1996.
- [Man 91] S. Manetti, "New approaches to automatic symbolic analysis of electric circuits," *Proc. IEE*, Pt. G, pp. 22–28, February 1991.
- [Map 91] "Maple V language reference manual", Springer-Verlag, 1991.
- [May 87] K. Mayaram, J. Lee and C. Hu, "A model for the electric field in lightly doped drain structures," *IEEE Trans. Electron Devices*, Vol. ED-34, No. 7, pp. 1509–1518, July 1987.
- [Mc And 96] C. McAndrew and L. Nagel, "Early effect modeling in SPICE," *IEEE J. Solid-State Circuits*, Vol. 31, No. 1, pp. 136–138, January 1996.
- [Maes 84] W. Maes, K. De Meyer and L. Dupas, "DC characterization of MOS transistors, SPICE model level 3," *Internal Report Dept. of Electronics, Katholieke Universiteit Leuven, Belgium*, December 1984.
- [Mas 53] S. Mason, "Feedback theory some properties of signal flow graphs," *Proc. IRE*, pp. 1144–1156, September 1953.
- [Matt 96] M. Mattausch, U. Feldman, A. Rahm, M. Bollu and D. Savignac, "Unified complete MOSFET model for analysis of digital and analog circuits," *IEEE Trans. Computer-Aided Design*, Vol. 15, No. 1, pp. 1–7, January 1996.
- [May 57] W. Mayeda and S. Seshu, "Topological formulas for network functions," *Engineering Experimentation Station Bulletin 446*, University of Illinois, Urbana, 1957.
- [Merck 72] G. Merckel, J. Borel and N. Cupcea, "An accurate large-signal MOS transistor model for use in computer-aided design," *IEEE Trans. Electron Devices*, Vol. ED-19, No. 5, pp. 681–690, May 1972.
- [Mey 72] R. Meyer, M. Shensa, and R. Eschenbach, "Cross modulation and intermodulation in amplifiers at high frequencies" *IEEE J. Solid-State Circuits*, Vol. SC-7, No. 1, pp. 16–23, February 1972.
- [Mey 86] R. Meyer, "Intermodulation in high-frequency bipolar transistor integrated-circuit mixers," *IEEE J. Solid-State Circuits*, Vol. SC-21, No. 4, pp. 534–537, August 1986.
- [Mey 94] R. Meyer and W. Mack, "A 1 GHz BiCMOS RF front-end IC," IEEE J. Solid-State Circuits, Vol. 29, No. 3, pp. 350–355, March 1994.

- [Mey 97] R. Meyer and W. Mack, "A 2.5 GHz BiCMOS transceiver for wireless LAN," Proc. ISSCC 1997, 1997, pp. 310–311.
- [Moon 91] B.-J. Moon, C.-K. Park, K. Lee and M. Shur, "New short-channel n-MOSFE' current-voltage model in strong inversion and unified parameter-extraction method, *IEEE Trans. Electron Devices*, Vol. 38, No. 3, pp. 592–602, March 1991.
- [Mull 86] R. Muller and T. Kamins, *Device electronics for integrated circuits*. John Wiley & Sons, 1986.
- [Nar 67] S. Narayanan, "Transistor distortion analysis using Volterra series representation, *The Bell System Technical J.*, Vol. 46, pp. 991–1024, May/June 1967.
- [Nar 70] S. Narayanan, "Application of Volterra series to intermodulation distortion analysi of transistor feedback amplifiers," *IEEE Trans. Circuit Theory*, Vol. CT-17, No. 4 pp. 518–527, November 1970.
- [Nar 73] S. Narayanan and H. Poon, "An analysis of distortion in bipolar transistors using integral charge control model and Volterra series," *IEEE Trans. Circuit Theory*, Vol. CT 20, No. 4, pp. 341–351, July 1973.
- [Neb 95] G. Nebel, U. Kleine and H.J. Pfleiderer, "Symbolic pole/zero calculation using SANTAFE," *IEEE J. Solid-State Circuits*, Vol. SC-30, No. 7, pp. 752–761, July 1995
- [Ngu 90] N. Nguyen and R. Meyer, "Si IC-compatible inductors and LC passive filters," *IEEl J. Solid-State Circuits*, Vol. 25, pp. 1028–1031, August 1990.
- [Num 92] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical recipes in C.* Cambridge University Press, 1992.
- [Opt 89] F. Op 't Eynde, P. Wambacq and W. Sansen, "On the relationship between the CMRI or PSRR and the second harmonic distortion of differential-input amplifiers," *IEEE Solid-State Circuits*, Vol. 24, No. 6, pp. 1740–1744, December 1989.
- [Park 92] H. Park, P. K. Ko and C. Hu, "A non-quasi-static MOSFET model for SPICE A analysis," *IEEE Trans. Computer-Aided Design*, Vol. 11, No. 10, pp. 1247–125 October 1992.
- [Pat 90] W. Patterson and F. Shoucair, "Harmonic suppression in unbalanced analog MOSFE circuit topologies using body signals," *Proc. ISCAS*, 1990, pp. 1151–1154.
- [Pel 89] M. Pelgrom, A. Duinmaijer and A. Welbers, "Matching properties of MOS transitors," *IEEE J. Solid-State Circuits*, Vol. 24, No. 5, pp. 1433–1439, 1989.
- [Pool 84] D. Poole and D. Kwong, "Two-dimensional analysis model of threshold voltage short-channel MOSFET's," *IEEE Electron Device Letters*, Vol. EDL-5, p. 443, 198

- [Poon 69] H. Poon and H. Gummel, "Modeling of emitter capacitance," *IEEE Proc. (Letters)*, Vol. 57, pp. 2181–2182, December 1969.
- [Poon 72] H. Poon, "Modeling of bipolar transistor using integral charge-control model with application to third-order distortion studies," *IEEE Trans. Electron Devices*, Vol. ED-19, No. 6, pp. 719–731, June 1972.
- [Poor 80] T. Poorter and J. Satter, "A DC model for an MOS-transistor in the saturation region," *Solid-State Electronics*, Vol. 22, pp. 701–717, 1979.
- [Popa 72] A. Popa, "An injection level dependent theory of the MOS transistor in saturation," *IEEE Trans. Electron Devices*, Vol. ED-19, pp. 774–781, June 1972.
- [Pow 92] J. Power and W. Lane, "An enhanced SPICE MOSFET model suitable for analog applications," *IEEE Trans. Computer-Aided Design*, Vol. 11, No. 11, pp. 1418–1425, November 1992.
- [Ross 76] P. Rossel, H. Martinot and G. Vassilieff, "Accurate two-sections model for MOS transistors in saturation," *Solid-State Electronics*, Vol. 19, pp. 51–56, January 1976.
- [Royc 89] J. Roychowdhury, "SPICE 3 distortion analysis," Memo. No. UCB/ERL M89/48, Electron. Res. Lab., Univ. of California, Berkeley, April 1989.
- [Rud 78] M. Rudko and D. Weiner, "Volterra systems with random inputs: a formalized approach," *IEEE Trans. Communications*, Vol. COM-26, No. 2, pp. 217–225, February 1978.
- [Sab 79] A. Sabnis and J. Clemens, "Characterization of the electron mobility in the inverted < 100 > Si surface," *IEDM Technical Digest*, pp. 18–21, Washington, 1979.
- [Sale 82] A. A. M. Saleh, "Matrix analysis of mildly nonlinear, multiple-input, multiple-output systems with memory," *The Bell System Tech. J.*, Vol. 61, No. 9, pp. 2221–2243, November 1982.
- [Sann 80] P. Sannuti and N.N. Puri, "Symbolic network analysis an algebraic formulation," *IEEE Trans. Circuit Theory*, Vol. CAS-27, No. 8, pp. 679–687, August 1980.
- [Sans 72] W. Sansen, "Optimum design of integrated variable-gain amplifiers," Ph.D. Dissertation Univ. of California, Berkeley, 1972.
- [Sata 90] H. Satake and T. Hamasaki, "Low-temperature (77K) BJT model with temperature dependences on the injected condition and base resistance," *IEEE Trans. Electron Devices*, Vol. 35, pp. 1055–1062, 1988.
- [Sato 96] H. Sato et al., "A 1.9 GHz single chip IF transceiver for digital cordless phones," *IEEE J. Solid-State Circuits*, Vol. 31, No. 12, pp. 1974–1980, December 1996.

- [Sche 80] M. Schetzen, The Volterra and Wiener theories of nonlinear systems. J. Wiley & Sons 1980.
- [Schw 83] S. Schwarz and S. Russek, "Semi-empirical equations for electron velocity in silicon part II—MOS inversion layer," *IEEE Trans. Electron Devices*, Vol. ED-30, No. 12 pp. 1634–1639, December 1983.
- [Seven 91] J. Sevenhans, A. Vanwelsenaers, J. Wenin and J. Baro, "An integrated Si bipola transceiver for a zero IF 900 MHz GSM digital mobile radio front-end of a han portable phone," *Proc. CICC*, 1991, pp. 7.7.1–7.7.4.
- [Sheu 84] B. Sheu and P. Ko, "An analytical model for intrinsic capacitances of short-channed MOSFETs," *Proc. IEDM*, 1984, pp. 300–303.
- [Sheu 87] B. Sheu, D. Sharfetter, P.-K. Ko and M.-C. Jeng, "BSIM: Berkeley short-channed IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, Vol. SC-22, No. 4 pp. 558–566, August 1987.
- [Shich 68] H. Shichman and D. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE J. Solid-State Circuits*, Vol. SC-3, No. 3, pp. 285, September 1968.
- [Shou 92] F. Shoucair, "A semi-empirical model of the MOSFET's small-signal drain condutance in saturation for analog circuit design," *IEEE Trans. Electron Devices*, Vol. 3 No. 5, pp. 1246–1248, May 1992.
- [Shou 93] F. Shoucair and W. Patterson, "Analysis and modeling of nonlinearities in VL MOSFET's including substrate effects," *IEEE Trans. Electron Devices*, Vol. 4 NO. 10, pp. 1760–1767, October 1993.
- [Skel 80] S. Skelboe, "Computation of the periodic steady-state response of nonlinear network by extrapolation methods," *IEEE Trans. Circuits and Systems*, Vol. CAS-27, No. pp. 161–175, March 1980.
- [Smit 94] D. Smit, M. Koen and A. Witulski, "Evolution of high-speed operational ampliful architectures," *IEEE J. Solid-State Circuits*, Vol. 29, No. 10, pp. 1166–1179, Octob 1994.
- [Solo 74] J. Solomon, "The monolithic opamp: a tutorial study," *IEEE J. Solid-State Circul* Vol. SC-9, No. 6, pp. 314–332, December 1974.
- [Somm 93] R. Sommer, E. Hennig, G. Dr'oge and E.H. Horneber, "Equation-based symbol approximation by matrix reduction with quantitative error prediction," *Alta Freques Rivista di Elettronica*, Vol. 5, No. 6, pp. 29–37, November 1993.

- [Sodi 84] C. Sodini, P. K. Ko and J. Moll, "The effect of high fields on MOS device and circuit performance," *IEEE Trans. Electron Devices*, Vol. ED-31, No. 10, pp. 1386–1393, October 1984.
- [Stet 95] T. Stetzler, I. Post, J. Havens and M. Koyama, "A 2.7-4.5 V single chip GSM transceiver RF integrated circuit," *IEEE J. Solid-State Circuits*, Vol. 30, No. 12, pp. 1421–1429, December 1995.
- [Sze 85] S.M. Sze, Semiconductor devices physics and technology. John Wiley & Sons, 1985.
- [Taft 91] R. Taft and J. Plummer, "An eight-terminal Kelvin-tapped bipolar transistor for extracting parasitic series resistances," *IEEE Trans. Electron Devices*, Vol. 38, No. 9, pp. 2139–2154, September 1991.
- [Tay 79] G. Taylor, "The effect of two-dimensional charge sharing on the above-threshold characteristics of short-channel devices," *Solid-State Electronics*, Vol. 22, pp. 701–717, 1979.
- [Tho 68] L. Thomas, "Eliminating broadband distortion in transistor amplifiers," *Bell Syst. Tech. J.*, Vol. 47, pp. 315–342, March 1968.
- [Toh 88] K.-Y. Toh, P.-K. Ko and R. Meyer, "An engineering model for short-channel MOS devices," *IEEE J. Solid-State Circuits*, Vol. 23, No. 4, pp. 950–958, August 1988.
- [Toy 79] T. Toyabe and S. Asai, "Analytical models of threshold voltage and breakdown voltage of short-channel MOSFET's derived from two-dimensional analysis," *IEEE Trans. Electron Devices*, Vol. ED-26, pp. 453–461, 1973.
- [Tsiv 81] Y. Tsividis and D. Fraser, "Harmonic distortion in single-channel MOS integrated circuits," *IEEE J. Solid-State Circuits*, Vol. SC-16, No. 6, pp. 694–702, December 1981.
- [Tsiv 83] Y. Tsividis and G. Masetti, "Problems in precision modeling of the MOS transistor for analog applications," *IEEE Trans. Computer-Aided Design*, Vol. CAD-3, No. 1, pp. 72–79, January 1983.
- [Tsiv 88] Y. Tsividis, Operation and modeling of the MOS transistor. McGraw-Hill International Editions, 1988.
- [Tsiv 93a] Y. Tsividis and J. Voorman, *Integrated continuous-time filters*. New York: IEEE Press, 1993.
- [Tsiv 93b] Y. Tsividis and K. Suyama, "MOSFET modeling for analog circuit CAD: problems and prospects," *Proc. CICC*, 1993, pp. 14.1.1–14.1.6.
- [Tsiv 94] Y. Tsividis, "Integrated continuous-time filter design an overview," *IEEE J. Solid-State Circuits*, Vol. 29, No. 3, pp. 166–176, March 1994.

- [Turch 83] C. Turchetti, G. Masetti and Y. Tsividis, "On the small-signal behavior of the MO transistor in quasi-static operation," *Solid-State Electronics*, Vol. 26, pp. 941–949 1983.
- [VdEi 89] E. Van den Eijnde, "Steady-state analysis of strongly nonlinear circuits," PhD dissertation, Vrije Universiteit Brussel, 1989.
- [VdWal 83] J. Vandewalle, J. Rabaey, W. Vercruysse, and H. De Man, "Computer-aided distortion analysis of switched capacitor filters in the frequency domain," *IEEE J. Solia State Circuits*, Vol. SC-18, No. 3, pp. 324-333, June 1983.
- [VdWie 79] F. Vandewiele, "A long-channel MOSFET model," *Solid-State Electronics*, pp. 991-997, December 1979.
- [Vlach 83] J. Vlach and K. Singhal, Computer methods for circuit analysis and design. Val Nostrand Reinhold, 1983.
- [Voi 97] S. Voinigescu and M. Maliepaard, "5.8 GHz and 12.6 GHz Si bipolar MMICs," *Proc ISSCC 1997*, pp. 372–373.
- [Wamb 90] P. Wambacq, G. Gielen, and W. Sansen, "Symbolic simulation of harmonic distortion in analog integrated circuits with weak nonlinearities," *Proc. ISCAS*, May 1990, pp. 536–539.
- [Wamb 91a] P. Wambacq, G. Gielen and W. Sansen, "Interactive symbolic distortion analysis of analogue integrated circuits," *Proc. EDAC*, February 1991, pp. 484–488.
- [Wamb 91b] P. Wambacq, G. Gielen, J. Vanthienen, and W. Sansen, "A design tool for weakly nonlinear analog integrated circuits with multiple inputs," *Proc. CICC*, 1991 pp. 5.1.1-5.1.4.
- [Wamb 92] P. Wambacq, G. Gielen, and W. Sansen, "A cancellation-free algorithm for the symbolic analysis of large analog circuits," *Proc. ISCAS*, 1992, pp. 1157–1160.
- [Wamb 94a] P. Wambacq, F. Fernández, G. Gielen, W. Sansen and A. Rodríguez-Vázquez, "A gorithm for efficient symbolic analysis of large analogue circuits," *IEE Electron Letters*, Vol. 30, No. 14, pp. 1108–1109, July 1994.
- [Wamb 95] P. Wambacq, F. Fernández, G. Gielen, W. Sansen, and A. Rodríguez-Vázquez, "ficient symbolic computation of approximated small-signal characteristics," *IEEE Solid-State Circuits*, Vol. 30, No. 3, pp. 327–330, March 1995.
- [Wamb 96] P. Wambacq, "Symbolic analysis of large and weakly nonlinear analog integration circuits," Ph.D. Dissertation Univ. of Leuven, Belgium, 1996.
- [Wamb 97] P. Wambacq, G. Gielen and W. Sansen, "Symbolic network analysis methods practical analog integrated circuits: a survey," accepted for publication in IEEE Trace.

  Circuits and Systems, Part II.

- [Wein 80] D. Weiner, J. Spina, Sinusoidal analysis and modeling of weakly nonlinear circuits. Van Nostrand Reinhold, 1980.
- [White 80] M. White, F. Van De Wiele and J.-P. Lambot, "High-accuracy MOS models for computer-aided design," *IEEE Trans. Electron Devices*, Vol. ED-27, No. 5, pp. 899–906, May 1980.
- [Whitt 69] R. Whittier and D. Tremere, "Current gain and cutoff frequency falloff at high currents," *IEEE Trans. Electron Devices*, Vol. ED-16, No. 1, pp. 39–57, January 1969.
- [Wier 89] G. Wierzba et al., "SSPICE—A symbolic SPICE program for linear active circuits," Proc. Midwest Symp. on Circuits and Systems, 1989.
- [Wolf 91] S. Wolfram, Mathematica, A System for Doing Mathematics by Computer. Wolfram, Massachusetts, 1991.
- [Wu 85] C.-Y. Wu and Y.-W. Daih, "An accurate mobility model for the I-V characteristics of n-channel enhancement-mode MOSFETs with single-channel boron implantation," *Solid-State Electronics*, Vol. 28, No. 12, pp. 1271–1278, 1985.
- [Yu 88] S. Yu, A. Franz, T. Mihran, "A physical parametric transistor model for CMOS circuit simulation," *IEEE Trans. Computer-Aided Design*, Vol. 7, pp. 1038–1052.
- [Yu 96] Q. Yu and C. Sechen, "A unified approach to the approximate symbolic analysis of large analog integrated circuits," *IEEE Trans. Circuits and Systems, Part I*, Vol. 43, No. 8, pp. 656–669, August 1996.
- [Yuan 88] J.-S. Yuan, J. Liou and W. Eisenstadt, "A physics-based current-dependent base resistance model for advanced bipolar transistors," *IEEE Trans. Electron Devices*, Vol. 35, No. 7, pp. 1055–1062, July 1988.

# Appendix A

# Useful trigonometric relationships

$$\exp(\pm jx) = \cos x \pm j \sin x$$

$$\cos x = \frac{1}{2} \left( \exp(jx) + \exp(-jx) \right)$$

$$\sin x = \frac{1}{2j} \left( \exp(jx) - \exp(-jx) \right)$$

$$\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$$

$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$$

$$\sin x \sin y = \frac{1}{2} \left( \cos(x - y) - \cos(x + y) \right)$$

$$\cos x \cos y = \frac{1}{2} \left( \cos(x+y) + \cos(x-y) \right)$$

$$\sin x \cos y = \frac{1}{2} \left( \sin(x+y) + \sin(x-y) \right)$$

$$sin2x = 2 \sin x \cos x$$

$$\cos 2x = 2\cos^2 x - 1 = 1 - 2\sin^2 x = \cos^2 x - \sin^2 y$$

$$\sin^2 x = \frac{1}{2} \left( 1 - \cos 2x \right)$$

$$\cos^2 x = \frac{1}{2} \left( 1 + \cos 2x \right)$$

$$\sin^3 x = \frac{1}{4} \left( 3\sin x - \sin 3x \right)$$

$$\cos^3 x = \frac{1}{4} \left( 3\cos x + \cos 3x \right)$$

$$\sin^4 x = \frac{1}{8} \left( \cos 4x - 4 \cos 2x + 3 \right)$$

$$\cos^4 x = \frac{1}{8} \left( \cos 4x + 4 \cos 2x + 3 \right)$$

# Appendix B

# **Basics of Volterra series**

#### **B.1** Introduction

The goal of this appendix is to provide the basics of Volterra series in order to give a better understanding of the material covered in Chapters 4 and 5. More details can be found in [Sche 80].

# **B.2** Volterra series representation of a system

The Volterra series operation can be viewed as a generalization of the theory of linear, first-order systems to weakly nonlinear systems. A nonlinear system is considered as the combination of different operators of different order.

The relationship between the input x(t) and the output y(t) of a nonlinear, time invariant system can, with certain restrictions, be expressed with the following series:

$$y(t) = \int_{-\infty}^{+\infty} h_1(\tau_1)x(t-\tau_1)d\tau_1$$

$$+ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_2(\tau_1, \tau_2)x(t-\tau_1)x(t-\tau_2)d\tau_1d\tau_2$$

$$+ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_3(\tau_1, \tau_2, \tau_3)x(t-\tau_1)x(t-\tau_2)x(t-\tau_3)d\tau_1d\tau_2d\tau_3$$

$$\dots + \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} h_n(\tau_1, \tau_2, \dots, \tau_n)x(t-\tau_1)x(t-\tau_2)\dots$$

$$\dots \cdot x(t-\tau_n)d\tau_1d\tau_2\dots d\tau_n + \dots$$
(B.1)

in which for  $n = 1, 2, \ldots$ ,

$$h_n(\tau_1, \tau_2, ..., \tau_n) = 0$$
 for any  $\tau_j < 0$ ,  $j = 1, 2, ..., n$  (B.2)

This series is called the *Volterra series* and the functions  $h_n(\tau_1, \tau_2, \dots, \tau_n)$  are called the *Volterra kernels* of the system. Another way of expressing equation (B.1) is

$$y(t) = \mathbf{H_1}[x(t)] + \mathbf{H_2}[x(t)] + \mathbf{H_3}[x(t)] + \dots + \mathbf{H_n}[x(t)] + \dots$$
(B.3)

in which

$$\mathbf{H}_n[x(t)] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} h_n(\tau_1, \tau_2, \dots, \tau_n) x(t - \tau_1) x(t - \tau_2) \cdots x(t - \tau_n) d\tau_1 d\tau_2 \dots d\tau_n$$
(B.4)

In this representation the symbol  $\mathbf{H}_n$ , which represents an integral, is called an *n*th-order *Volterra* operator. A schematic representation of equation (B.3) is shown in Figure B.1. Clearly, a non-

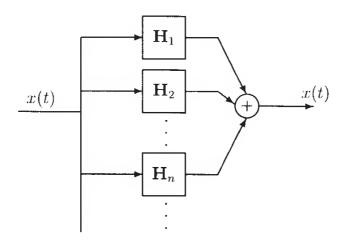


Figure B.1: Volterra series representation of a system.

linear system is considered as the combination of different operators of different order.

In the next sections, the individual terms  $\mathbf{H}_n$  are discussed using the concept of *n*th-order operators. The form in equation (B.4) of the Volterra kernel is used to calculate the response to sinusoidal excitations. In order to clarify the results, the second-order Volterra operator is considered before the generalization to the *n*th order.

The derivations are limited to systems with one input port. The extension to multiple-input systems, however, is straightforward [Chua 79a, Chua 79b].

# **B.3** Second-order Volterra systems

In this section the second-order Volterra operator, which has been qualitatively discussed in the previous section, is defined mathematically. The second-order operator gives rise to the second term in the Volterra series. It is shown how such an operator can be characterized by starting from the concept of a general second-order operator. Once the second-order Volterra operator is known, it can be used to study the response of the second-order system to a sinusoidal input.

#### **B.3.1** The second-order operator

A linear operator has been defined as one for which the response to a linear combination of signals is the same linear combination of the response to each input signal individually. Extending this concept to a second-order operator, one can define the latter as an operator for which the response to a linear combination of signals is a *bilinear* operation on the individual input signals. This means that, if the input is given by

$$x(t) = \sum_{i=1}^{N} c_i x_i(t)$$
(B.5)

in which the coefficients  $c_i$  are arbitrary complex constants, then the output as a result of the second-order operator is given by

$$y_{2}(t) = \mathbf{T}_{2}[x(t)]$$

$$= \mathbf{T}_{2} \left[ \sum_{i=1}^{N} c_{i} x_{i}(t) \right]$$

$$= \sum_{i_{1}=1}^{N} \sum_{i_{2}=1}^{N} \mathbf{T}_{2} \{ c_{i_{1}} x_{i_{1}}(t), c_{i_{2}} x_{i_{2}}(t) \}$$

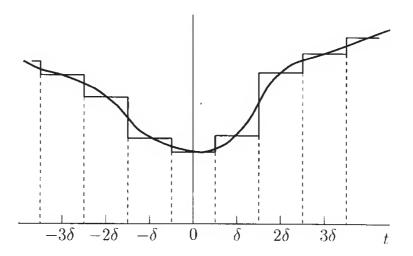
$$= \sum_{i_{1}=1}^{N} \sum_{i_{2}=1}^{N} c_{i} c_{i_{2}} \mathbf{T}_{2} \{ x_{i_{1}}(t), x_{i_{2}}(t) \}$$
(B.6)

The operator  $T_2\{\cdot\}$  is called a *bilinear operator* since  $T_2\{x,y\}$  is linear in x for a given y and also linear in y for a given x. Also,  $T_2\{x,x\} = T_2[x]$ . The response of a second-order system to a linear combination of input signals thus involves the sum of product operations on the input signals taken two at a time. This is the mathematical representation of the intuitive interpretation of a second-order nonlinearity as discussed in Chapter 4: a second-order nonlinearity combines two signals at its input. These signals can be identical.

## **B.3.2** The second-order Volterra operator

The operator  $T_2$  defined for second-order time-invariant systems is called a **second-order Volters operator**, and is denoted by  $H_2$ . Such a Volterra operator performs a transformation from the input signal to the output signal. This transformation is characterized now.

In order to obtain a general system characterization, a basic waveform must be chosen which all input waveforms can be represented as the linear combination given in equation (B.A.) A basic waveform to choose is obtained by observing that all waveforms of interest can be proximated by a staircase curve as shown in Figure B.2. Such a staircase curve can be express as the sum of displaced rectangles. The waveform corresponding to one such rectangle can



*Figure B.2: The staircase approximation*  $x_{\delta}(t)$  *of* x(t)*.* 

defined in terms of the function  $u_{\delta}(t)$ , which is given by :

$$u_{\delta}(t) = \begin{cases} \frac{1}{\delta} & \text{for } |t| < \frac{1}{2}\delta \\ 0 & \text{elsewhere} \end{cases}$$
 (B.7)

Note that for  $u_{\delta}(t)$ 

$$\int_{-\infty}^{+\infty} u_{\delta}(t)dt = 1 \tag{B.8}$$

Now the staircase approximation  $x_{\delta}$  for x(t), shown in Figure B.2, can be written as a sum:

$$x_{\delta}(t) = \sum_{i=-\infty}^{+\infty} \delta \ x(i\delta) \ u_{\delta}(t - k\delta)$$
 (B.9)

This form resembles to the linear combination of equation (B.5).

The staircase approximation to x(t) becomes better as  $\delta$  becomes smaller. In the limit, when  $\delta$  approaches zero, then  $x_{\delta}(t)$  goes to x(t). In order to calculate the response  $y_2(t)$  to x(t), the response  $y_{\delta}$  to  $x_{\delta}$  will be first calculated. Then  $y_2(t)$  is calculated from  $y_{\delta}(t)$  by letting  $\delta$  go to zero.

Using the second-order bilinear time-invariant operator  $\mathbf{H}_2$ , the output  $y_{\delta}(t)$  of a second-order system excited by the staircase waveform  $x_{\delta}$  is given by

$$y_{\delta}(t) = \mathbf{H}_{2}[x_{\delta}(t)]$$

$$= \sum_{i_{1}=-\infty}^{+\infty} \sum_{i_{2}=-\infty}^{+\infty} \mathbf{H}_{2} \{\delta x(i_{1}\delta)u_{\delta}(t-i_{1}\delta), \delta x(i_{2}\delta)u_{\delta}(t-i_{2}\delta)\}$$

$$= \sum_{i_{1}=-\infty}^{+\infty} \sum_{i_{2}=-\infty}^{+\infty} \delta x(i_{1}\delta)\delta x(i_{2}\delta)\mathbf{H}_{2} \{u_{\delta}(t-i_{1}\delta), u_{\delta}(t-i_{2}\delta)\}$$
(B.10)

To each bilinear operation in equation (B.10) a time function corresponds, denoted as  $h_2$ :

$$h_2(t - i_1\delta, t - i_2\delta, \delta) = \mathbf{H}_2\{u_\delta(t - i_1\delta), u_\delta(t - i_2\delta)\}$$
(B.11)

This time function is a function only of the distance in time between the two pulses and of  $\delta$ , the pulse width. The position of the pulses only determines the position of the time function on the time axis, since the system under consideration is time-invariant. In terms of equation (B.11), the response  $y_{\delta}(t)$  can be written as:

$$y_{\delta}(t) = \sum_{i_1 = -\infty}^{+\infty} \sum_{i_2 = -\infty}^{+\infty} \delta x(i_1 \delta) \delta x(i_2 \delta) h_2(t - i_1 \delta, t - i_2 \delta, \delta)$$
 (B.12)

When  $\delta$  goes to zero, then  $y_{\delta}(t)$  approaches  $y_{2}(t)$ , and the double summation can be rewritten as a double integral:

$$y_2(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x(\sigma_1)x(\sigma_2)h_2(t - \sigma_1, t - \sigma_2)d\sigma_1d\sigma_2$$
 (B.13)

In this equation, the time function  $h_2(\cdot)$  is given by

$$h_2(t - \sigma_1, t - \sigma_2) = \lim_{\delta \to 0} h_2(t - \sigma_1, t - \sigma_2, \delta)$$
  
=  $\mathbf{H}_2 \{ u_0(t - \sigma_1), u_0(t - \sigma_2) \}$  (B.14)

in which the function  $u_0(t)$  is the unit impulse, which is the limit of  $u_{\delta}(t)$  as  $\delta$  approaches zero.

Equation (B.13) gives the functional representation of the second-order Volterra kernel. It is seen to be a two-dimensional convolution, which can be even better recognized by putting  $\tau_1 = t - \sigma_1$  and  $\tau_2 = t - \sigma_2$ :

$$y_{2}(t) = \mathbf{H}_{2}[x(t)]$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2})x(t - \tau_{1})x(t - \tau_{2})d\tau_{1}d\tau_{2}$$
(B.15)

The function  $h_2(\tau_1, \tau_2)$  is called the **second-order Volterra kernel** of the Volterra operator  $\mathbf{H_2}$ . In [Sche 80] it is shown that the second-order Volterra kernel can be considered as the response of a new second-order system derived from the original one, to a set of two impulses, applied at a different time. This corresponds to the interpretation of the first term of the Volterra series which is a convolution of the input with the linear system's impulse response.

The output of the second-order system is thus a two-dimensional convolution of the input with the second-order Volterra kernel. This Volterra kernel is independent of the input signal. Hence the two-dimensional convolution integral allows to compute the output of a second-order system to any kind of input signal. In this work, the input is mostly a sinusoid or a sum of sinusoids, but the convolution integral can also be used to find the output as a result of a Gaussian input the represents white noise [Bedr 71, Rud 78, Sche 80] or any other input [Sche 80].

#### **B.3.3** Second-order kernel symmetrization

A second-order kernel is symmetric if  $h_2(\tau_1, \tau_2) = h_2(\tau_2, \tau_1)$ . In many analyses with Volterra operators, operations have to be performed that involve a reordering of the  $\tau$ 's in the kernel  $h_2(\tau_1, \tau_2)$ . The analyses are greatly simplified if only symmetric kernels need to be considered, for then the specific order of  $\tau$ 's is not important. Fortunately a procedure exists by which any asymmetric kernel can be symmetrized so that it is possible to consider only symmetric kernels without any loss of generality. Suppose that an asymmetric form  $h_2(\tau_1, \tau_2)$  is known. Then the symmetric form  $h_2(\tau_1, \tau_2)$  is given by

$$h_2(\tau_1, \tau_2) = \frac{1}{2} [h_2^*(\tau_1, \tau_2) + h_2^*(\tau_1, \tau_2)]$$
 (B.16)

Moreover, this symmetric form is unique [Sche 80].

#### **B.4** The second-order kernel transform

Apart from a time-domain description, Volterra kernels can also be represented in the frequency domain. This is an extension of a transfer function for linear systems. Just as for these systems, a Laplace and a Fourier transform can be defined.

## **B.4.1** The two-dimensional Fourier and Laplace transform

The two-dimensional Fourier transform is obtained by taking the Fourier transform of a function of two variables  $f_2(\tau_1, \tau_2)$ . In order to avoid convergence problems in the subsequent integrals, it is required that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |f_2(\tau_1, \tau_2)| d\tau_1 d\tau_2 < \infty \tag{B.17}$$

Suppose that  $\tau_2$  is held constant in  $f_2(\tau_1, \tau_2)$ , then the (one-dimensional) Fourier transform, with respect to  $\tau_1$  is given by

$$F_1(j\omega_1, \tau_2) = \int_{-\infty}^{+\infty} f(\tau_1, \tau_2) e^{-j\omega_1 \tau_1} d\tau_1$$
 (B.18)

When in this intermediate function the variable  $\omega_1$  is kept constant, then the one-dimensional Fourier transform of  $F_1(j\omega_1, \tau_2)$  is given by

$$F_2(j\omega_1, j\omega_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_2(\tau_1, \tau_2) e^{-j(\omega_1 \tau_1 + \omega_2 \tau_2)} d\tau_1 d\tau_2$$
 (B.19)

The function  $F_2(j\omega_1, j\omega_2)$  is called the **two-dimensional Fourier transform** of the function  $f_2(\tau_1, \tau_2)$ . The two-dimensional inverse transform of  $F_2(j\omega_1, j\omega_2)$  can be obtained in a similar manner which yields

$$f_2(\tau_1, \tau_2) = \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F_2(j\omega_1, j\omega_2) e^{j(\omega_1 \tau_1 + \omega_2 \tau_2)} d\omega_1 d\omega_2$$
 (B.20)

The equations (B.19) and (B.20) establish the two-dimensional Fourier transform pair  $f_2(\tau_1, \tau_2)$  and  $F_2(j\omega_1, j\omega_2)$ .

An interesting property which can be proven easily is that

$$H_2(-j\omega_1, -j\omega_2) = H_2^*(j\omega_1, j\omega_2)$$
(B.21)

If the kernel  $h_2$  is symmetric, then the Fourier transform is symmetric as well:

$$H_2(j\omega_1, j\omega_2) = H_2(j\omega_2, j\omega_1)$$
(B.22)

Similarly to the Fourier transform, the **two-dimensional Laplace transform** of a kernel  $f_2(\tau_1, \tau_2)$  is defined as

$$F_2(s_1, s_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_2(\tau_1, \tau_2) e^{-(s_1 \tau_1 + s_2 \tau_2)} d\tau_1 d\tau_2$$
 (B.23)

in which  $s_1 = \sigma_1 + j\omega_1$  and  $s_2 = \sigma_2 + j\omega_2$ . The Fourier transform can be derived from the Laplace transform by letting  $\sigma_1 = \sigma_2 = 0$ . This requires that the Fourier transform of  $f_2(\tau_1, \tau_2)$  exists so that the region of absolute convergence for  $F_2(s_1, s_2)$  includes the  $\omega_1$  and the  $\omega_2$  axes.

## **B.4.2** Sinusoidal response of a second-order Volterra system

Assume that a second-order system is excited by a sinusoidal input

$$x(t) = A_x \cos \omega_x t \tag{B.24}$$

With  $x_a(t) = (A_x/2) \exp(j\omega_x t)$  and  $x_b(t) = (A_x/2) \exp(-j\omega_x t)$  the input can be rewritten as

$$x(t) = x_a(t) + x_b(t) \tag{B.25}$$

, The response of the second-order system to this input is given by

$$y_2(t) = \mathbf{H}_2[x_a(t)] + \mathbf{H}_2[x_b(t)] + \mathbf{H}_2\{x_a(t), x_b(t)\} + \mathbf{H}_2\{x_b(t), x_a(t)\}$$
(B.26)

Since the kernel corresponding to  $\mathbf{H}_2$  can be considered symmetric, one could let  $\mathbf{H}_2$   $\{x_a(t), x_b(t)\}$  =  $\mathbf{H}_2$   $\{x_b(t), x_a(t)\}$  but this is not necessary for the derivations here. The first term of equation (B.26) is now written in terms of the second-order Volterra kernel using the two-dimensional convolution:

$$\mathbf{H}_{2}[x_{a}(t)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) x_{a}(t - \tau_{1}) x_{a}(t - \tau_{2}) d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) e^{j\omega_{x}(t - \tau_{1})} e^{j\omega_{x}(t - \tau_{2})} d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} e^{j2\omega_{x}t} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) e^{-j\omega_{x}\tau_{1}} e^{-j\omega_{x}\tau_{2}} d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} H_{2}(j\omega_{x}, j\omega_{x}) e^{j2\omega_{x}t}$$
(B.27)

The second term of equation (B.26) can be calculated similarly, which yields

$$\mathbf{H}_2[x_b(t)] = \frac{A_x^2}{4} H_2(-j\omega_x, -j\omega_x) e^{-j2\omega_x t}$$
(B.28)

The third term of equation (B.26) is given by

$$\mathbf{H}_{2} \left\{ x_{a}(t), x_{b}(t) \right\} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) x_{a}(t - \tau_{1}) x_{b}(t - \tau_{2}) d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) e^{j\omega_{x}(t - \tau_{1})} e^{-j\omega_{x}(t - \tau_{2})} d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{2}(\tau_{1}, \tau_{2}) e^{-j\omega_{x}\tau_{1}} e^{+j\omega_{x}\tau_{2}} d\tau_{1} d\tau_{2} 
= \frac{A_{x}^{2}}{4} H_{2}(j\omega_{x}, -j\omega_{x})$$
(B.29)

and similarly for the fourth term,

$$\mathbf{H}_{2}\left\{x_{b}(t), x_{a}(t)\right\} = \frac{A_{x}^{2}}{4} H_{2}(-j\omega_{x}, j\omega_{x})$$
(B.30)

Adding the four terms of equation (B.26) yields the output y(t)

$$y_2(t) = \left(\frac{A_x}{2}\right)^2 H_2(j\omega_x, j\omega_x) e^{j2\omega_x t} + \left(\frac{A_x}{2}\right)^2 H_2(-j\omega_x, -j\omega_x) e^{-j2\omega_x t}$$

$$\left(\frac{A_x}{2}\right)^2 H_2(j\omega_x, -j\omega_x) + \left(\frac{A_x}{2}\right)^2 H_2(-j\omega_x, j\omega_x)$$
(B.31)

The first two terms are complex conjugates of each other and so are the last two ones. This reduces the output to the final result

$$y_2(t) = 2\left(\frac{A_x}{2}\right)^2 \operatorname{Re}\left(H_2(j\omega_x, j\omega_x)e^{j2\omega_x t}\right) + 2\left(\frac{A_x}{2}\right)^2 \operatorname{Re}\left(H_2(j\omega_x, -j\omega_x)\right)$$
(B.32)

Equation (B.32) shows that the response of a second-order system to a sinusoidal input is a DC term and a sinusoid at the double frequency.

If the second-order kernel is symmetric, which can be assumed without any loss of generality, then  $H_2(j\omega_x, -j\omega_x)$  turns out to be a real number, since the interchange of the arguments does not change the value and at the same time yields the complex conjugate.

Note that if only a complex exponential input is applied, then the output is given by the first term only of equation (B.31).

## B.4.3 Response of a second-order system to a sum of two sinusoids

We now calculate the response of the second-order system to the following input:

$$x(t) = A_x \cos \omega_x t + A_y \cos \omega_y t$$

$$= \frac{A_x}{2} e^{j\omega_x t} + \frac{A_x}{2} e^{-j\omega_x t} + \frac{A_y}{2} e^{j\omega_y t} + \frac{A_y}{2} e^{-j\omega_y t}$$

$$= x_a(t) + x_b(t) + x_c(t) + x_d(t)$$
(B.33)

The second-order kernel is assumed to be symmetric. In that case the response is

$$y_{2}(t) = \mathbf{H}_{2}[x_{d}(t)] = \mathbf{H}_{2}[x_{a}(t) + x_{b}(t) + x_{c}(t) + x_{d}(t)]$$

$$= \mathbf{H}_{2}[x_{a}(t)] + \mathbf{H}_{2}[x_{b}(t)] + \mathbf{H}_{2}[x_{c}(t)] + \mathbf{H}_{2}[x_{d}(t)]$$

$$+2\mathbf{H}_{2}\{x_{a}(t), x_{b}(t)\} + 2\mathbf{H}_{2}\{x_{a}(t), x_{c}(t)\}$$

$$+2\mathbf{H}_{2}\{x_{a}(t), x_{d}(t)\} + 2\mathbf{H}_{2}\{x_{b}(t), x_{c}(t)\}$$

$$+2\mathbf{H}_{2}\{x_{b}(t), x_{d}(t)\} + 2\mathbf{H}_{2}\{x_{c}(t), x_{d}(t)\}$$
(B.34)

The expression of the response in terms of the second-order nonlinear transfer function can be computed similarly to the previous section. The result is

$$y_{2}(t) = \frac{A_{x}^{2}}{2} \operatorname{Re} \left( H_{2}(j\omega_{x}, j\omega_{x}) e^{j2\omega_{x}t} \right) + \frac{A_{x}^{2}}{2} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{x}) \right)$$

$$+ \frac{A_{y}^{2}}{2} \operatorname{Re} \left( H_{2}(j\omega_{y}, j\omega_{y}) e^{j2\omega_{y}t} \right) + \frac{A_{y}^{2}}{2} \operatorname{Re} \left( H_{2}(j\omega_{y}, -j\omega_{y}) \right)$$

$$+ A_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, j\omega_{y}) e^{j(\omega_{x} + \omega_{y})t} \right)$$

$$+ A_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

$$+ B_{x} A_{y} \operatorname{Re} \left( H_{2}(j\omega_{x}, -j\omega_{y}) e^{j(\omega_{x} - \omega_{y})t} \right)$$

The first four terms in equation (B.35) correspond to the ones obtained when one single sinusoid is applied. The last two terms are intermodulation products.

# **B.5** Higher-order Volterra systems

The theory presented in the previous section for the second order can be extended to the *pth* order. The derivations again start from a general *p*-linear operator.

## **B.5.1** The pth-order operator

Before a generalization to order p, the third-order operator is discussed first. This operator denoted as  $T_3$  is defined as one for which the response to a linear combination of signals is a trilinear operation on the individual input signals. This means that, if the input is again a linear combination of signals given by

$$x(t) = \sum_{i=1}^{N} c_i x_i(t) \tag{B.36}$$

then the output as a result of the third-order operator  $T_3$  is given by

$$y_{3}(t) = \mathbf{T}_{3}[x(t)]$$

$$= \mathbf{T}_{3} \left[ \sum_{i=1}^{N} c_{i}x_{i}(t) \right]$$

$$= \sum_{i_{1}=1}^{N} \sum_{i_{2}=1}^{N} \sum_{i_{3}=1}^{N} \mathbf{T}_{3} \{c_{i_{1}}x_{i_{1}}(t), c_{i_{2}}x_{i_{2}}(t), c_{i_{3}}x_{i_{3}}(t) \}$$

$$= \sum_{i_{1}=1}^{N} \sum_{i_{2}=1}^{N} \sum_{i_{3}=1}^{N} c_{i_{1}}c_{i_{2}}c_{i_{3}}\mathbf{T}_{3} \{x_{i_{1}}(t), x_{i_{2}}(t), x_{i_{3}}(t) \}$$

$$= \sum_{i_{1}=1}^{N} \sum_{i_{2}=1}^{N} \sum_{i_{3}=1}^{N} c_{i_{1}}c_{i_{2}}c_{i_{3}}\mathbf{T}_{3} \{x_{i_{1}}(t), x_{i_{2}}(t), x_{i_{3}}(t) \}$$
(B.37)

The operator  $T_3\{\cdot\}$  is called a *trilinear operator* since  $T_3\{x,y\}$  is linear with respect to each argument when the other two are held constant. Also,  $T_3\{x,x,x\} = T_3[x]$ . The response of a second-order system to a linear combination of input signals thus involves the sum of product operations on the input signals taken three at a time.

In a similar manner, a **p-linear operator**,  $T_p$  can be defined. The response  $y_p$  of that operator to the input given by equation (B.36) is

$$y_{p}(t) = \mathbf{T}_{p}[x(t)]$$

$$= \mathbf{T}_{p} \left[ \sum_{i=1}^{N} c_{i} x_{i}(t) \right]$$

$$= \sum_{i_{1}=1}^{N} \cdots \sum_{i_{p}=1}^{N} \mathbf{T}_{p} \{ c_{i_{1}} x_{i_{1}}(t), \dots, c_{i_{p}} x_{i_{p}}(t) \}$$

$$= \sum_{i_{1}=1}^{N} \cdots \sum_{i_{p}=1}^{N} c_{i_{1}} \cdots c_{i_{p}} \mathbf{T}_{p} \{ x_{i_{1}}(t), \dots, x_{i_{p}}(t) \}$$
(B.38)

The p-linear operator  $T_p$  is linear with respect to every argument when all the others are kept constant. Also, if  $x_1 = x_2 = \cdots = x_p = x$  then  $T_p\{x, \ldots, x\} = T_p[x]$ .

# B.5.2 The pth-order Volterra operator

The operator  $T_p$  defined for a pth-order time-invariant system is called a pth-order Volterra operator, and is denoted by  $H_p$ . The functional representation of this operator can be found in a similar fashion as for the second-order case. First the response is computed of the pth-order system to a staircase approximation of the input x(t) given in equation (B.9). When the width of the pulses in this approximation goes to zero then the response approaches the real output  $y_p(t)$ ,

which is found to be

$$y_{p}(t) = \mathbf{H}_{p}[x(t)]$$

$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} h_{p}(\tau_{1}, \tau_{2}, \dots, \tau_{p}) x(t - \tau_{1}) x(t - \tau_{2}) \cdots x(t - \tau_{p}) d\tau_{1} d\tau_{2} \dots d\tau_{p}$$
(B.39)

This is also the pth term of the Volterra series representation of a general nonlinear system. The function  $h_p(\tau_1, \tau_2, \ldots, \tau_p)$  is called the **pth-order Volterra kernel** of the Volterra operator  $\mathbf{H}_p$ . The output of a pth-order system is a p-dimensional convolution of the input with the Volterra kernel of order p.

For a third-order system, the output given in equation (B.39) reduces to

$$y_{3}(t) = \mathbf{H}_{3}[x(t)]$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h_{3}(\tau_{1}, \tau_{2}, \tau_{3}) x(t - \tau_{1}) x(t - \tau_{2}) x(t - \tau_{3}) d\tau_{1} d\tau_{2} d\tau_{3}$$
 (B.40)

## B.5.3 pth-order kernel symmetrization

In section B.3.3 it was pointed out that the second-order Volterra kernel of a second-order system can be considered to be symmetric without any loss of generality. Also, if an asymmetric form is known, then the symmetric form can be easily found. These results can be extended to order p. Suppose that an asymmetric form  $h_p^*(\tau_1, \ldots, \tau_p)$  is known. Then the symmetric form  $h_p(\tau_1, \ldots, \tau_p)$  is given by

$$h_p(\tau_1, \dots, \tau_p) = \frac{1}{p!} \sum_{\substack{\text{all possible} \\ \text{rearrangements} \\ \text{of } \tau_1, \dots, \tau_p}} h_p^*(\tau_1, \dots, \tau_p)$$
(B.41)

Note that for p = 2 this equation reduces to equation (B.16).

Since a symmetrization is always possible, it is assumed in the rest of this appendix as well as in the calculations with Volterra kernels in other chapters that the Volterra kernels are symmetric.

## **B.5.4** The p-dimensional Laplace and Fourier transforms

This section generalizes the results from Section B.4.1 for functions of p variables,  $f_p(\tau_1, \ldots, \tau_p)$  for which

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} |f_p(\tau_1, \dots, \tau_p)| d\tau_1 \cdots d\tau_p < \infty$$
(B.42)

The multidimensional Laplace transform  $F_p(s_1, \ldots, s_p)$  is obtained by repeatedly taking the Laplace transform with respect to one variable. In this way, one obtains

$$F_p(s_1,\ldots,s_p) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_p(\tau_1,\ldots,\tau_p) e^{-(s_1\tau_1+\ldots s_2\tau_p)} d\tau_1 \cdots d\tau_p$$
 (B.43)

in which  $s_i = \sigma_i + j\omega_i$  for i = 1...p. The multidimensional Fourier transform is defined similarly by making all  $\sigma_i$  equal to zero in equation (B.43).

Similar to the second-order operator, the response of a pth-order operator to a sinusoidal signal or a sum of sinusoids can be computed. For example, the response of a third- and fourth-order operator to a sinusoidal excitation  $A_x \cos(\omega_x t)$  is given by

$$y_3(t) = \mathbf{H}_3 \left[ A_x \cos \omega_x t \right]$$

$$= \frac{A_x^3}{4} \operatorname{Re}(H_3(j\omega_x, j\omega_x, j\omega_x) e^{j3\omega_x t})$$

$$= \frac{3A_x^3}{4} \operatorname{Re}(H_3(j\omega_x, j\omega_x, -j\omega_x) e^{j\omega_x t})$$
(B.44)

and

$$y_{4}(t) = \mathbf{H}_{4} \left[ A_{x} \cos \omega_{x} t \right]$$

$$= \frac{A_{x}^{4}}{8} \operatorname{Re}(H_{4}(j\omega_{x}, j\omega_{x}, j\omega_{x}, j\omega_{x}) e^{j4\omega_{x}t})$$

$$= \frac{A_{x}^{4}}{2} \operatorname{Re}(H_{4}(j\omega_{x}, j\omega_{x}, j\omega_{x}, -j\omega_{x}) e^{j2\omega_{x}t})$$

$$= \frac{3A_{x}^{4}}{4} \operatorname{Re}(H_{4}(j\omega_{x}, j\omega_{x}, -j\omega_{x}, -j\omega_{x}))$$
(B.45)

# **Appendix C**

# Derivation of the method for the direct computation of nonlinear responses

In this appendix a method is derived to compute harmonics and intermodulation products without making use of Volterra series. The method computes responses to at most two sinusoidal input signals that are applied simultaneously and possibly at different input ports. Assuming that the frequency of the two input sinusoids is  $\omega_1$  and  $\omega_2$ , respectively, then a node voltage in the circuit under consideration consists of components at frequencies  $|\pm m\omega_1 \pm n\omega_2|$ , where m and n are positive integers. The complex amplitude or phasor of such component is written as  $V_{i,\pm m,\pm n}$  where i corresponds to the numbering of the nodes in the circuit. Below it is explained how such phasor is computed.

# **C.1** Setup of basic equations

The network shown in Figure C.1 contains linear circuit elements among which the capacitor C and nonlinear circuit elements among which a nonlinear transconductance. The derivation given here will focus only on that transconductance. For the other basic nonlinearities that have been defined in Section 3.2, the derivations are completely similar. The network is excited by two different current sources  $i_{in1}$  and  $i_{in2}$ , applied at two different ports. For voltage source excitations, the derivation is similar.

The nonlinear transconductance is controlled by the voltage difference between nodes i and j. The nonlinear relationship between the controlled current  $i_1$  and the controlling voltage is given by:

$$i_{1} = f(v_{i}(t) - v_{j}(t))$$

$$= g_{m}(v_{i}(t) - v_{j}(t)) + K_{2}g_{m}(v_{i}(t) - v_{j}(t))^{2} + K_{3}g_{m}(v_{i}(t) - v_{j}(t))^{3} + \dots$$
(C.1)

The sinusoidal excitations are given by:

$$i_{in1}(t) = \text{Re}\left(I_{in1}e^{j\omega_1 t}\right) = \frac{1}{2}I_{in1}e^{j\omega_1 t} + \frac{1}{2}I_{in1}^*e^{-j\omega_1 t}$$
 (C.2)

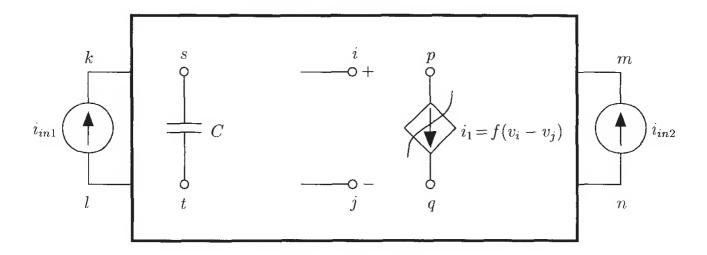


Figure C.1: A general nonlinear network.

and

$$i_{in2}(t) = \frac{1}{2} I_{in2} e^{j\omega_2 t} + \frac{1}{2} I_{in2}^* e^{-j\omega_2 t}$$
 (C.3)

Under steady-state conditions, the AC-part of every node voltage  $v_x(t)$  can be regarded as being composed of a sum of harmonics and intermodulation products:

$$v_x(t) = \sum_{m=0}^{+\infty} \sum_{n=0}^{+\infty} \text{Re}\left(V_{x,m,n} e^{j(m\omega_1 + n\omega_2)t}\right) = \frac{1}{2} \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{x,m,n} e^{j(m\omega_1 + n\omega_2)t}$$
(C.4)

in which  $V_{x,m,n}$  is the phasor of the voltage at node x at the frequency  $m\omega_1 + n\omega_2$  and  $V_{x,m,n} = V_{x,-m,-n}^*$ .

Applying Kirchoff's current law in the given network yields on the different nodes:

$$(k) \dots + \frac{-I_{in1}}{2}e^{j\omega_1 t} + \frac{-I_{in1}^*}{2}e^{-j\omega_1 t} \dots = 0$$
(C.5)

$$(l) \dots + \frac{I_{in1}}{2}e^{j\omega_1 t} + \frac{I_{in1}^*}{2}e^{-j\omega_1 t} \dots = 0$$
(C.6)

$$(m) \dots + \frac{-I_{in2}}{2} e^{j\omega_2 t} + \frac{-I_{in2}^*}{2} e^{-j\omega_2 t} \dots = 0$$
 (C.7)

$$\widehat{n} \dots + \frac{I_{in2}}{2} e^{j\omega_2 t} + \frac{I_{in2}^*}{2} e^{-j\omega_2 t} \dots = 0$$
(C.8)

$$(p) \dots + \frac{1}{2} g_m \left( \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{i,m,n} e^{j(m\omega_1 + n\omega_2)t} - \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{j,m,n} e^{j(m\omega_1 + n\omega_2)t} \right)$$

$$+\frac{1}{4}K_{2g_{m}}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{i,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{j,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)^{2}$$

$$+\frac{1}{8}K_{3g_{m}}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{x,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{j,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)^{3}$$

$$+\ldots=0 \tag{C.9}$$

$$q)\ldots-\frac{1}{2}g_{m}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{i,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{j,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)^{2}$$

$$-\frac{1}{4}K_{2g_{m}}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{i,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{j,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)^{2}$$

$$-\frac{1}{8}K_{3g_{m}}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{x,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{j,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)^{3}$$

$$+\ldots=0 \tag{C.10}$$

$$s)\ldots+C\frac{d}{dt}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{x,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{t,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)$$

$$+\ldots=0 \tag{C.11}$$

$$t)\ldots+C\frac{d}{dt}\left(\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{t,m,n}e^{j(m\omega_{1}+n\omega_{2})t}-\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}V_{x,m,n}e^{j(m\omega_{1}+n\omega_{2})t}\right)$$

$$+\ldots=0 \tag{C.11}$$

The equations for nodes p and q contain terms with second and third powers. Expansion of these powers yields for the equation of node p:

$$\dots + \frac{1}{2} g_{m} \cdot \left( \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{i,m,n} e^{j(m\omega_{1}+n\omega_{2})t} - \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} V_{j,m,n} e^{j(m\omega_{1}+n\omega_{2})t} \right)$$

$$+ \frac{1}{4} K_{2} g_{m} \cdot \left( \sum_{m_{a}=-\infty}^{+\infty} \sum_{n_{a}=-\infty}^{+\infty} \sum_{m_{b}=-\infty}^{+\infty} \sum_{n_{b}=-\infty}^{+\infty} (V_{i,m_{a},n_{a}} V_{i,m_{b},n_{b}} + V_{j,m_{a},n_{a}} V_{j,m_{b},n_{b}} - 2V_{i,m_{a},n_{a}} V_{j,m_{b},n_{b}}) e^{j((m_{a}+m_{b})\omega_{1}+(n_{a}+n_{b})\omega_{2})t} \right)$$

$$+ \frac{1}{8} K_{3} g_{m} \cdot \left( \sum_{m_{a}=-\infty}^{+\infty} \sum_{n_{a}=-\infty}^{+\infty} \sum_{m_{b}=-\infty}^{+\infty} \sum_{n_{b}=-\infty}^{+\infty} \sum_{m_{c}=-\infty}^{+\infty} (V_{i,m_{a},n_{a}} V_{i,m_{b},n_{b}} V_{i,m_{c},n_{c}} - 3 V_{i,m_{a},n_{a}} V_{i,m_{b},n_{b}} V_{j,m_{c},n_{c}} + 3 V_{i,m_{a},n_{a}} V_{j,m_{b},n_{b}} V_{j,m_{c},n_{c}} - V_{j,m_{a},n_{a}} V_{j,m_{b},n_{b}} V_{j,m_{c},n_{c}} \right) e^{j((m_{a}+m_{b}+m_{c})\omega_{1}+(n_{a}+n_{b}+n_{c})\omega_{2})t} + \dots = 0$$
(C.13)

A similar expansion of the powers can be performed in the equation for node q.

It is seen that these higher-order powers cause effects of order higher than one. For example, in the expansion of the second power, a term of order k is caused by a combination of the indices  $m_a, m_b, n_a, n_b$  for which

$$|m_a| + |m_b| + |n_a| + |n_b| = k$$
 (C.14)

After a complete expansion of the powers in the Kirchoff equations, one obtains a set of equations that consist of nothing else but sums of complex exponentials of the form  $\exp(j\omega_k t)$  , each multiplied with a complex coefficient. These equations must be valid for any value of the time variable t. This is only possible when the coefficients of every complex exponential are equal in both sides of every Kirchoff equation. This equality is now specified for every frequency of interest.

Before doing so, an important simplifying assumption is made. It is presumed that a harmonic or an intermodulation product of a certain order k is only determined by the combination of responses of order lower than k. For example, the contributions of order three, five, ... to the response at  $\omega_1$  are neglected. The error introduced by these neglections is small in low-distortion applications.

#### C.2First-order responses

The first complex powers to be considered are  $\exp(\pm j\omega_1 t)$  and  $\exp(\pm j\omega_2 t)$ . The retrieval of the coefficients of  $\exp(+j\omega_1 t)$  from the Kirchoff current equations (C.5) through (C.12) yields:

$$(k) \rightarrow \ldots + \frac{-I_{in1}}{2} \ldots = 0$$
 (C.15)

$$\overbrace{n} \to \dots = 0$$
(C.18)

$$(p) \rightarrow \dots + \frac{1}{2} g_m (V_{i,1,0} - V_{j,1,0})$$

$$+ \frac{1}{4} K_{2} g_m \text{ (higher-order terms)}$$

$$+ \frac{1}{8} K_{3} g_m \text{ (higher-order terms)}$$

$$+ \dots = 0 \tag{C.19}$$

$$q$$
  $\rightarrow ... - \frac{1}{2}g_m (V_{i,1,0} - V_{j,1,0})$ 

$$-\frac{1}{4}K_{2}g_{m} \; (\text{higher-order terms})$$
 
$$-\frac{1}{8}K_{3}g_{m} \; (\text{higher-order terms})$$
 
$$+\ldots=0 \tag{C.20}$$

$$s \to \dots + \frac{1}{2} j \omega_1 C \left( V_{s,1,0} - V_{t,1,0} \right) + \dots = 0$$
 (C.21)

$$t \rightarrow \dots + \frac{1}{2} j \omega_1 C \left( V_{t,1,0} - V_{s,1,0} \right) + \dots = 0$$
 (C.22)

When the terms of order higher than one are neglected, then the equations (C.15) through (C.22) reduce to the network equations in the frequency domain of the linearized equivalent of the given circuit, excited by a current source  $i_{in1} = I_{in1} \exp(j\omega_1 t)$ . From these equations one finds the complex amplitudes  $V_{x,1,0}$ . The circuit that corresponds to the equations (C.15) through (C.22) is depicted in Figure C.2.

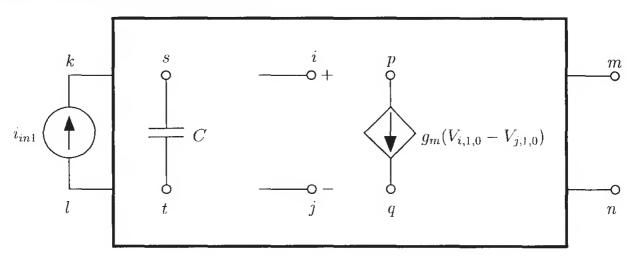


Figure C.2: The linearized equivalent of the network of Figure C.1 used to find the first-order responses to  $i_{in1}$ .

A similar procedure for the coefficients of  $\exp(-j\omega_1 t)$  again yields a set of linear equations. These correspond now to the network equations of the linearized circuit, excited by a current source  $i_{in1} = I_{in1} \exp(-j\omega_1 t)$ . From these equations the responses  $V_{x,-1,0}$  are computed. One will find, of course, that  $V_{x,-1,0}$  is the complex conjugate of  $V_{x,1,0}$ .

In a similar way, the coefficients of  $\exp(\pm j\omega_2 t)$  in the Kirchoff equations yield a set of linear equations from which the complex amplitudes  $V_{x,0,\pm 1}$  are found. These represent the response of the linearized circuit to the current source  $i_{in2} = I_{in2} \exp(\pm j\omega_2 t)$ .

## C.3 Second-order responses

The retrieval of coefficients of  $\exp(\pm j\omega_1 t)$  and  $\exp(\pm j\omega_2 t)$  resulted in the knowledge of the phasors at the frequencies  $\pm \omega_1$  and  $\pm \omega_2$ . These results are now used for the calculation of

second-order harmonics and intermodulation products.

Suppose that the response at the sum or difference frequency  $\omega_1 \pm \omega_2$  must be known. This second-order response can be found by taking the coefficients of  $\exp(j\omega_1 \pm \omega_2)$  in the expanded Kirchoff equations (C.5) through (C.12). This yields:

$$(C.23)$$

$$(C.24)$$

$$(m) \rightarrow \dots = 0$$
 (C.25)

$$(n) \rightarrow \dots = 0$$
 (C.26)

$$\begin{array}{l} (p) \rightarrow \ldots + \frac{1}{2} g_m \left( V_{i,1,\pm 1} - V_{j,1,\pm 1} \right) \\ + \frac{1}{4} K_{2} g_m \left( V_{i,1,0} V_{i,0,\pm 1} + V_{i,0,\pm 1} V_{i,1,0} + V_{j,1,0} V_{j,0,\pm 1} + V_{j,0,\pm 1} V_{j,1,0} \right. \\ \left. - 2 V_{i,1,0} V_{j,0,\pm 1} - 2 V_{i,0,\pm 1} V_{j,1,0} + \text{higher-order terms} \right) \\ + \frac{1}{8} K_{3} g_m \left( \text{higher-order terms} \right) + \ldots \\ + \ldots = 0 \end{array} \tag{C.27}$$

$$\begin{array}{l} \boxed{q} \rightarrow \ldots - \frac{1}{2} g_m \left( V_{i,1,\pm 1} - V_{j,1,\pm 1} \right) \\ - \frac{1}{4} K_{2} g_m \left( V_{i,1,0} V_{i,0,\pm 1} + V_{i,0,\pm 1} V_{i,1,0} + V_{j,1,0} V_{j,0,\pm 1} + V_{j,0,\pm 1} V_{j,1,0} \right. \\ - 2 V_{i,1,0} V_{j,0,\pm 1} - 2 V_{i,0,\pm 1} V_{j,1,0} + \text{higher-order terms}) \\ - \frac{1}{8} K_{3} g_m \left( \text{higher-order terms} \right) - \ldots \\ + \ldots = 0 \end{array} \tag{C.28}$$

$$s \to \dots + \frac{1}{2} j(\omega_1 \pm \omega_2) C(V_{s,1,\pm 1} - V_{t,1,\pm 1}) + \dots = 0$$
 (C.29)

$$t \to \dots + \frac{1}{2} j(\omega_1 \pm \omega_2) C(V_{t,1,\pm 1} - V_{s,1,\pm 1}) + \dots = 0$$
 (C.30)

Neglecting the higher-order terms and introducing the auxiliary variable  $i_{NL2gm}$ 

$$i_{NL2g_m} = K_{2g_m} (V_{i,1,0} - V_{j,1,0}) (V_{i,0,\pm 1} - V_{j,0,\pm 1})$$
 (C.31)

one obtains

$$(C.32)$$

$$(C.33)$$

$$(m) \rightarrow \dots = 0$$
 (C.34)

$$\overbrace{n} \to \dots = 0$$
(C.35)

$$(p) \to \dots + g_m (V_{i,1,\pm 1} - V_{j,1,\pm 1}) + \dots = -i_{NL2g_m}$$
 (C.36)

$$(q) \rightarrow \dots - g_m (V_{j,1,\pm 1} - V_{i,1,\pm 1}) + \dots = +i_{NL2g_m}$$
 (C.37)

$$s \to \dots + \frac{1}{2} j(\omega_1 \pm \omega_2) C(V_{s,1,\pm 1} - V_{t,1,\pm 1}) + \dots = 0$$
 (C.38)

$$\underbrace{t} \to \dots + \frac{1}{2} j(\omega_1 \pm \omega_2) C(V_{t,1,\pm 1} - V_{s,1,\pm 1}) + \dots = 0$$
 (C.39)

Comparing these equations with the equations (C.15) through (C.22), it is clear that the equations above correspond to the same linear network, with the unknown node voltages now being  $V_{x,1,\pm 1}$  instead of  $V_{x,1,0}$ , and the frequency being  $\omega_1 \pm \omega_2$ . Also, the external excitations are removed. Instead a new excitation  $i_{NL2g_m}$  is applied. This is a fictitious current source, applied in parallel with the transconductance  $g_m$ , which is the linearized equivalent of the nonlinear transconductance. The orientation of the source is from the positive node of the nonlinearity to the negative one. The circuit that corresponds to the equations (C.32) through (C.39) is depicted in Figure C.3.

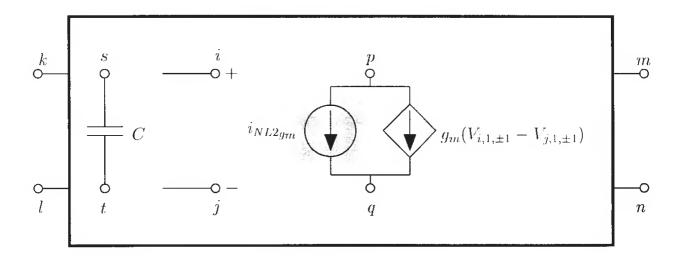


Figure C.3: The linearized equivalent of the network of Figure C.1 excited by  $i_{NL2gm}$  to find the responses at  $\omega_1 \pm \omega_2$ .

The value of the current source is determined by the second-order nonlinearity coefficient of the nonlinearity under consideration and by the first-order response of its controlling voltage(s) due to the excitations  $i_{in1}$  and  $i_{in2}$ , respectively.

When a circuit contains other nonlinearities than the transconductance we considered, then additional current sources need to be considered, one for every nonlinearity. Their value depends upon the kind of basic nonlinearity. Table 5.5 lists the value of the current source for the different basic nonlinearities defined in Section 3.2.

The complex amplitudes of voltages  $v_X$  in the circuit at frequency  $-\omega_1 \mp \omega_2$ , denoted as  $V_{x,-1,\pm 1}$ , are computed from a set of equations which is generated by taking the coefficients.

of  $\exp(j(-\omega_1\mp\omega_2)t)$  in the Kirchoff equations. When this set is solved, one will find that  $V_{x,-1,\mp 1}=V_{x,1,\pm 1}^*$ .

Other second-order responses are determined in a similar fashion: by considering the coefficients of the corresponding complex exponential, a set of equations is found that corresponds to the linearized equivalent of the circuit of Figure C.1 excited by nonlinear current sources. The values of the sources that have to be applied for the computation of the harmonic at  $2\omega_1$ , are given in the rightmost column of Table 5.5.

## C.4 Higher-order responses

For the computation of responses of order three, the same procedure leads to the solution of the same linearized network that is excited now with nonlinear current sources of order three. Their value depends upon first- and second-order responses for the controlling voltages as well as on the second- and third-order nonlinearity coefficients. For example, the expressions for the nonlinear current sources for the computation of the third-order intermodulation product at  $2\omega_1 \pm \omega_2$  and the third harmonic of  $\omega_2$  are given in Table 5.6 and 5.7, respectively.

# **Appendix D**

# Nonlinearity coefficients for the description of the Early effect

In this appendix, it is shown how the derivatives of the collector current of a bipolar transistor in its forward active operating region with respect to the collector-emitter voltage can be obtained. First, the derivatives with respect to  $v_{BC}$  will be computed. Form these values, the derivatives with respect to  $v_{CF}$  can be obtained easily.

According to the Gummel-Poon model [Getr 76, Anto 88], the collector current in the forward active region can be written as

$$i_C = \frac{I_S Q_{B0}}{Q_B} \exp\left(\frac{v_{BE}}{n_F V_t}\right) \tag{D.1}$$

in which  $Q_B$  is the majority base charge and  $Q_{B0}$  is the majority base charge at  $v_{BC}=0V$ .

For the derivation of the dependence on  $v_{CE}$ , low injection is considered here. Under low-injection conditions and neglecting the "late effect", which is the Early effect under reverse operation,  $Q_B$  is given by

$$Q_B = Q_{B0} + Q_C \tag{D.2}$$

in which  $Q_C$  is the charge due to the acceptor atoms in the base (for an npn transistor) which is uncovered as the collector-base depletion region retreats from its invasion into the neutral base region. Hence  $Q_C$  is a function of  $v_{BC}$  only and the collector current can be written as the product of a function of  $v_{BE}$  only and a function of  $v_{BC}$  only:

$$i_C = \frac{I_S Q_{B0} f(v_{BE})}{Q_{B0} + Q_C(v_{BC})} = I_S Q_{B0} f(v_{BE}) g(v_{BC})$$
 (D.3)

with

$$f(v_{BE}) = \exp\left(\frac{v_{BE}}{n_F V_t}\right) \tag{D.4}$$

$$g(v_{BC}) = \frac{1}{Q_{B0} + Q_C(v_{BC})}$$
(D.5)

The first-order derivative of the collector current with respect to  $v_{BC}$  is given by

$$\frac{\partial i_C}{\partial v_{BC}} = -\frac{I_S Q_{B0} f(v_{BE})}{(Q_{B0} + Q_C(v_{BC}))^{-2}} \cdot \frac{dQ_C(v_{BC})}{dv_{BC}}$$
(D.6)

The derivative of the charge  $Q_C(v_{BC})$  with respect to  $v_{BC}$  is nothing else but the junction capacitance  $C_{\mu}(v_{BC})$ , such that the derivative of the collector current becomes

$$\frac{\partial i_C}{\partial v_{BC}} = -\frac{I_S Q_{B0} f(v_{BE})}{(Q_{B0} + Q_C(v_{BC}))^{-2}} C_{\mu}(v_{BC})$$
 (D.7)

If  $V_{BC} = 0$  then this derivative becomes

$$\frac{\partial i_C}{\partial v_{BC}}\Big|_{V_{BC}=0V} = -\frac{I_S f(v_{BE})C_{\mu}(0)}{Q_{B0}}$$
 (D.8)

since  $Q_C(0) = 0$ . The ratio  $Q_{B0}/C_{\mu}(0)$  is defined as the Early voltage  $V_{AF}$ :

$$V_{AF} = \frac{Q_{B0}}{C_{\mu}(0)} = \frac{Q_{B0}}{C_{jc}} \tag{D.9}$$

in which  $C_{jc}$  is a shorthand notation for  $C_{\mu}(0)$ .

Using the Early voltage, the derivative of the collector current at  $V_{BC}=0V$  now becomes

$$\frac{\partial i_C}{\partial v_{BC}}\Big|_{V_{BC}=0V} = -\frac{I_S f(v_{BE})}{V_{AE}} \tag{D.10}$$

The numerator in the right-hand side is the collector current in absence of the Early effect. In the currently used SPICE models, the value of the derivative of the collector current with respect to  $v_{BC}$  for values of  $v_{BC}$  other than zero, is taken equal to the value for  $v_{BC} = 0V$ , as given in equation (D.10). This is not correct, as shown for example with the measurements presented in Chapter 9. Also, a constant value of the first derivative implies that all higher-order derivatives are zero, which is an oversimplification. A more accurate expression for the derivatives of the collector current with respect to  $v_{BC}$  is now derived.

As stated above, the junction capacitance  $C_{\mu}(v_{BC})$  is given by

$$C_{\mu}(v_{BC}) = \frac{dQ_C(v_{BC})}{dv_{BC}} \tag{D.11}$$

from which  $Q_C(v_{BC})$  is found to be

$$Q_C(v_{BC}) = \int_0^{v_{BC}} C_{\mu}(v) dv$$
 (D.12)

Using equation (3.107) we obtain

$$Q_C(v_{BC}) = \frac{C_{jc} \left( V_{jc} - v_{BC} \right)^{-mjc+1} V_{jc}^{mjc}}{mjc - 1} - \frac{V_{jc}}{mjc - 1}$$
(D.13)

Using this expression, a more accurate expression for  $\partial i_C/\partial v_{BC}$  is obtained:

$$\frac{\partial i_{C}}{\partial v_{BC}} = \frac{I_{S} f(v_{BE}) V_{AF} (mjc - 1)^{2} (V_{jc} - v_{BC})^{mjc} V_{jc}^{mjc}}{\left(\left(V_{AF} (V_{jc} - v_{BC})^{mjc}\right) \cdot (mjc - 1) + V_{jc}^{mjc+1} - V_{jc}^{mjc} v_{BC} - V_{jc} (V_{jc} - v_{BC})^{mjc}\right)^{2}}$$
(D.14)

The higher-order derivatives can now be derived from equation (D.14).

For the computation of cross-derivatives of  $i_C$  with respect to both  $v_{BC}$  and  $v_{BE}$ , the collector current is considered as a product of a function of  $v_{BE}$  and  $v_{BC}$ , as formulated in equation (D.3). Then one finds for the derivatives of the collector current of order one to three

$$\frac{\partial^n i_C}{\partial v_{BE}^n} = I_S Q_{B0} \cdot \frac{d^n f(v_{BE})}{dv_{BE}^n} \cdot g(v_{BC}) \qquad (n = 1, 2, 3)$$
(D.15)

$$\frac{\partial^n i_C}{\partial v_{BC}^n} = I_S Q_{B0} \cdot \frac{d^n g(v_{BC})}{dv_{BC}^n} \cdot f(v_{BE}) \qquad (n = 1, 2, 3)$$
(D.16)

$$\frac{\partial^2 i_C}{\partial v_{BE} \partial v_{BC}} = I_S Q_{B0} \cdot \frac{df(v_{BE})}{dv_{BE}} \cdot \frac{dg(v_{BC})}{v_{BC}}$$
(D.17)

$$\frac{\partial^3 i_C}{\partial v_{BE} \partial v_{BC}^2} = I_S Q_{B0} \cdot \frac{df(v_{BE})}{dv_{BE}} \cdot \frac{d^2 g(v_{BC})}{v_{BC}^2}$$
(D.18)

$$\frac{\partial^3 i_C}{\partial v_{BE}^2 \partial v_{BC}} = I_S Q_{B0} \cdot \frac{d^2 f(v_{BE})}{dv_{BE}^2} \cdot \frac{dg(v_{BC})}{v_{BC}}$$
(D.19)

(D.20)

In the model described in Chapter 3, the Early effect is modeled by making the collector current dependent on  $v_{CE}$  instead of  $v_{BC}$ . Then derivatives with respect to  $v_{CE}$  can be obtained from derivatives with respect to  $v_{BC}$  by making the substitution

$$v_{BC} = v_{BE} - v_{CE} \tag{D.21}$$

The derivative of the collector current with respect to  $v_{CE}$  is then found to be

$$\frac{\partial i_C}{\partial v_{CE}} = \frac{\partial i_C}{\partial v_{BC}} \cdot \frac{\partial v_{BC}}{\partial v_{CE}} \tag{D.22}$$

Using equation (D.21), this can be rewritten as

$$\frac{\partial i_C}{\partial v_{CE}} = -\frac{\partial i_C}{\partial v_{BC}} \tag{D.23}$$

For the second-order derivative we find

$$\frac{\partial^2 i_C}{\partial v_{CE}^2} = \frac{\partial^2 i_C}{\partial v_{BC}^2} \tag{D.24}$$

and so on.

Finally, it should be noted that with the substitution  $v_{BC} = v_{BE} - v_{CE}$ , the function g of equation (D.3) now becomes a function of both  $v_{BE}$  and  $v_{CE}$ . However, when the derivatives of the collector current are computed, the derivatives of g with respect to  $v_{BE}$  are neglected compared to the derivatives of f with respect to  $v_{BE}$ . In other words, the dependence of the Early effect on  $v_{BE}$  is neglected compared to the exponential relationship between  $v_{BE}$  and the collector current without the Early effect. The error made with this assumption is negligible.

#### **Appendix E**

# Relation between source-referred and bulk-referred nonlinearity coefficients of a MOS transistor

In Chapter 7, the drain current of a MOS transistor has been expressed both in terms of voltages referred to the source and in terms of voltages to the bulk. With the first reference system, the voltages that determine the drain current  $i_D$  are  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ . This reference system is the most widely used. On the other hand, when voltages are referred to the bulk, then the drain current model clearly shows the symmetry of source and drain. With this reference system the drain current is expressed in terms of the voltages  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ . Of course, both reference systems are equivalent. However, the nonlinearity coefficients in the first reference system are proportional to derivatives with respect to  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ , whereas in the other reference system the nonlinearity coefficients are proportional to derivatives with respect to  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ . The nonlinearity coefficients for both reference systems are defined in Table 7.2 and Table 7.3, respectively. In this appendix the relationship between the two sets of nonlinearity coefficients is considered.

To this purpose, the voltages of the source reference system are first expressed in terms of the voltages of the bulk reference system:

$$v_{GS} = v_{GB} - v_{SB} \tag{E.1}$$

$$v_{DS} = v_{DB} - v_{SB} \tag{E.2}$$

$$v_{SB} = v_{SB} \tag{E.3}$$

The voltage  $v_{SB}$  is used in both reference systems.

Assume now that we have the drain current equation in terms of source-referred voltages:

$$i_D = i_D(v_{GS}, v_{DS}, v_{SB})$$
 (E.4)

We now compute the derivative of  $i_D$  with respect to  $v_{GB}$ :

$$\frac{\partial i_D}{\partial v_{GB}} = \frac{\partial i_D}{\partial v_{GS}} \cdot \frac{\partial v_{GS}}{\partial v_{GB}} \tag{E.5}$$

The derivative in the left-hand side is the coefficient  $g_{mg}$ , as seen in Table 7.3. The first derivative in the right-hand side is  $g_m$ . The second derivative is computed from equation (E.1). Clearly, this derivative is equal to one. Hence we find

$$g_{mg} = g_m \tag{E.6}$$

Consider now the second-order derivative with respect to  $v_{GB}$ . From equation (E.5) we find

$$\frac{\partial^2 i_D}{\partial v_{GB}^2} = \frac{\partial^2 i_D}{\partial v_{GS}^2} \cdot \left(\frac{\partial v_{GS}}{\partial v_{GB}}\right)^2 + \frac{\partial i_D}{\partial v_{GS}} \cdot \frac{\partial^2 v_{GS}}{\partial v_{GB}^2}$$
(E.7)

From equation (E.1) we know that the relationship between  $v_{GS}$  and  $v_{GB}$  is linear. Hence the second derivative of  $v_{GS}$  with respect to  $v_{GB}$  is zero, such that the second term in the right-hand side of equation (E.7) is zero. Consequently, we find

$$\frac{\partial^2 i_D}{\partial v_{GB}^2} = \frac{\partial^2 i_D}{\partial v_{GS}^2} \tag{E.8}$$

and hence

$$K_{2q_{mg}} = K_{2q_m} (E.9)$$

Similarly we find

$$K_{3g_{mq}} = K_{3g_m} (E.10)$$

We now consider the derivatives with respect to  $v_{DB}$  and to  $v_{DS}$ . In the same way as above we easily find

$$g_{md} = g_0 \tag{E.11}$$

$$K_{2g_{md}} = K_{2g_o} (E.12)$$

$$K_{3g_{md}} = K_{3g_o} (E.13)$$

Next the derivatives with respect to  $v_{SB}$  are considered. To this purpose we write the first derivative of the drain current  $i_D(v_{GS}, v_{DS}, v_{SB})$  with respect to  $v_{SB}$  in terms of the derivative of the drain current expressed as a function of  $v_{GB}$ ,  $v_{DB}$  and  $v_{SB}$ :

$$\frac{\partial i_{D}(v_{GS}, v_{DS}, v_{SB})}{\partial v_{SB}} = \frac{\partial i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{GB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} + \frac{\partial i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}}$$
(E.14)

The left-hand side of this equation is nothing else but  $-g_{mb}$ . The partial derivatives of the drain current in the right-hand side are found from Table 7.3. Using equations (E.1) and (E.2), we find

$$\frac{\partial v_{GB}}{\partial v_{SB}} = \frac{\partial v_{DB}}{\partial v_{SB}} = 1 \tag{E.15}$$

Substituting this information into equation (E.14) yields

$$g_{mb} = -g_{mg} - g_{md} + g_{ms} (E.16)$$

We now consider the second derivative of  $i_D$  with respect to  $v_{SB}$ . From equation (E.14) we find

$$\frac{\partial^{2} i_{D}(v_{GS}, v_{DS}, v_{SB})}{\partial v_{SB}^{2}} = \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{GB}^{2}} \cdot \left(\frac{\partial v_{GB}}{\partial v_{SB}}\right)^{2} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{GB}\partial v_{DB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{GB}\partial v_{SB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{DB}^{2}} \cdot \left(\frac{\partial v_{DB}}{\partial v_{SB}}\right)^{2} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{DB}\partial v_{GB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{DB}\partial v_{SB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}^{2}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{GB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{GB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{GB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{GB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}} \\
+ \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{GB}} \cdot \frac{\partial v_{GB}}{\partial v_{SB}} + \frac{\partial^{2} i_{D}(v_{GB}, v_{DB}, v_{SB})}{\partial v_{SB}\partial v_{DB}} \cdot \frac{\partial v_{DB}}{\partial v_{SB}}$$
(E.17)

Using equations (E.1) through (E.3) together with the definitions of the nonlinearity coefficients listed in Table 7.3 we find

$$K_{2g_{mb}} = -K_{2g_{mg}} - K_{2g_{mg}\&g_{md}} - K_{2g_{mg}\&g_{md}} - K_{2g_{ms}\&g_{md}} + K_{2g_{ms}}$$
(E.18)

Similarly we find for the third-order nonlinearity coefficient  $K_{3g_{mb}}$ 

$$K_{3g_{mb}} = -K_{3g_{mg}} - K_{3g_{mg}\&g_{md}} - K_{3g_{mg}\&g_{ms}} - K_{3g_{mg}\&2g_{md}} - K_{3g_{mg}\&2g_{ms}} - K_{3g_{md}} - K_{3g_{md}\&g_{ms}} - K_{3g_{md}\&2g_{ms}} + K_{3g_{ms}} - K_{3g_{md}\&g_{md}\&g_{ms}}$$
(E.19)

Finally we consider the cross-derivatives. Nonlinearity coefficients that are proportional to cross-derivatives with respect to  $v_{GB}$  and  $v_{DB}$  only, are identical to nonlinearity coefficients that are proportional to cross-derivatives with respect to  $v_{GS}$  and  $v_{DS}$  only. In other words:

$$K_{2_{q_{mq} \& q_{md}}} = K_{2_{q_m \& q_o}} \tag{E.20}$$

$$K_{3_{g_{mg}\&2g_{md}}} = K_{3_{g_{m}\&2g_{g}}} \tag{E.21}$$

$$K_{3_{2g_{mg}\&g_{md}}} = K_{3_{2g_{m}\&g_{o}}} \tag{E.22}$$

For nonlinearity coefficients corresponding to cross-derivatives that contain at least one differentiation with respect to  $v_{SB}$ , relationships are found similar to the ones given in equations (E.18) and (E.19). For example, for  $K_{2g_m\&g_{mb}}$  we find

$$K_{2g_{m}\&g_{mb}} = 2K_{2g_{mg}} + 2K_{2g_{mg}\&g_{md}} + 2K_{2g_{mg}\&g_{ms}} + 2K_{2g_{md}} + 2K_{2g_{md}\&g_{ms}} - 2K_{2g_{ms}}$$
(E.23)

#### **Appendix F**

## Derivatives of the drain current with an implicit saturation voltage

If the drain current of a MOS transistor in the triode region is modeled with an explicit expression as a function of the terminal voltages  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$ , then it depends on the form of this expression whether an explicit expression of  $v_{DSAT}$  can be obtained. If no such explicit expression can be obtained, then no explicit expression of the drain current in the saturation region can be obtained as a function of the terminal voltages. Nevertheless, in such situation it is still possible to obtain closed-form expressions for the derivatives of the drain current with respect to the terminal voltages. In this appendix it will be shown how expressions for these derivatives can be obtained.

The expressions of the derivatives will contain  $v_{GS}$ ,  $v_{SB}$  and  $v_{DSAT}$ . Since  $v_{DSAT}$  is not known explicitly, only implicitly, an exact closed-form expression for the derivatives in which  $v_{DSAT}$  is substituted by its exact expression, cannot be obtained. If such expression is required, then one can approximate  $v_{DSAT}$  by an explicit expression for  $v_{DSAT}$  from a simpler model for which an analytic expression for  $v_{DSAT}$  can be obtained. The accuracy can be very good.

In this appendix expressions for the derivatives of the drain current are computed with a drain current model that only takes into account mobility reduction and a depletion layer with a width that varies in the direction of the channel. For the inclusion of effects such as velocity saturation the same approach can be used.

The equation for the drain current of a MOS transistor in the triode region is given in equation (7.71). This equation can be written in the form

$$i_D = mobred(v_{GS}, v_{DS}, v_{SB}) \cdot large(v_{GS}, v_{DS}, v_{SB})$$
 (F.1)

in which the function large is given by

$$large(v_{GS}, v_{DS}, v_{SB}) = \frac{\mu_0 C'_{ox} W}{L} \left\{ (v_{GS} - V_{FB} - \phi) v_{DS} - \frac{1}{2} v_{DS}^2 - \frac{2}{3} \gamma \left[ (\phi + v_{SB} + v_{DS})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
 (F.2)

The function *mobred* is defined by equation (7.73) and is repeated here for convenience:

$$mobred(v_{GS}, v_{DS}, v_{SB}) = \frac{1}{1 + f_u}$$
 (F.3)

where  $f_{\mu}$  is a function of  $v_{GS}$ ,  $v_{DS}$  and  $v_{SB}$  as given by equation (7.69).

If channel-length modulation is neglected, then the expression of the drain current in saturation can be found from equation (F.1) by substituting  $v_{DS}$  with  $v_{DSAT}$ :

$$i_{DSAT} = mobred(v_{GS}, v_{DSAT}, v_{SB}) \cdot large(v_{GS}, v_{DSAT}, v_{SB})$$
 (F.4)

Before proceeding to the determination of the derivatives, we determine  $v_{DSAT}$ .

#### **F.1** Determination of $v_{DSAT}$

The value of  $v_{DSAT}$  is determined from the drain current expression in the triode region. using equation (7.48). Using equation (F.4) we obtain

$$\frac{\partial mobred}{\partial v_{DS}} \cdot large + mobred \cdot \frac{\partial large}{\partial v_{DS}} = 0$$
 (F.5)

This equation can only be solved analytically for very simple models of *mobred* and *large*. For the model of equations (F.2) and (F.3) this is not possible. Hence equation (F.5) must be solved by iteration.

For the computations in the following sections equation (F.5) is written in a more compact form:

$$F(v_{DSAT}(v_{GS}, v_{SB}), v_{GS}, v_{SB}) = 0 (F.6)$$

in which F is the left-hand side of equation (F.5).

#### F.2 First-order derivatives

Assume we want to compute the first-order derivative of the drain current with respect to  $v_{GS}$ . Differentiation of equation (F.4) with respect to  $v_{GS}$  yields

$$\frac{\partial i_{DSAT}}{\partial v_{GS}} = \frac{\partial mobred}{\partial v_{GS}} \cdot large + \frac{\partial mobred}{\partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial v_{GS}} \cdot large + mobred \cdot \frac{\partial large}{\partial v_{GS}} + mobred \cdot \frac{\partial large}{\partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial v_{GS}}$$
(F.7)

in which the arguments of large and mobred have been omitted for clarity.

It is seen that the derivative of the current contains the derivative of  $v_{DSAT}$  with respect to  $v_{GS}$ . This derivative must be determined starting from equation (F.6). Differentiation of F with respect to  $v_{GS}$  yields

$$\frac{\partial F}{\partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial v_{GS}} + \frac{\partial F}{\partial v_{GS}} = 0$$
 (F.8)

in which the arguments of F are omitted for clarity.

From equation (F.8) one obtains the first-order derivative of  $v_{DSAT}$  with respect to  $v_{GS}$ :

$$\frac{\partial v_{DSAT}}{\partial v_{GS}} = -\frac{\partial F}{\partial v_{GS}} / \frac{\partial F}{\partial v_{DSAT}} \tag{F.9}$$

Explicit expressions for the derivatives in the right-hand side can be obtained from the expression of F. Hence, an explicit expression for  $\partial v_{DSAT}/\partial v_{GS}$  can be determined. The expression obtained in this way can now be substituted into equation (F.7). In this way, all necessary elements are available to find an explicit expression for  $\partial i_{DSAT}/\partial v_{GS}$  in terms of  $v_{GS}$ ,  $v_{SB}$  and  $v_{DSAT}$ .

#### F.3 Higher-order derivatives

Higher-order derivatives of the drain current with respect to one of the terminal voltages will contain higher-order derivatives of  $v_{DSAT}$  with respect to the terminal voltages. These can be easily obtained by a differentiation of the first derivatives of  $v_{DSAT}$ . For example, the second-order derivative of  $v_{DSAT}$  with respect to  $v_{GS}$  can be obtained from a further differentiation of equation (F.8) with respect to  $v_{GS}$ :

$$\frac{\partial^2 F}{\partial v_{DSAT}^2} \left( \frac{\partial v_{DSAT}}{\partial v_{GS}} \right)^2 + \frac{\partial F}{\partial v_{DSAT}} \frac{\partial^2 v_{DSAT}}{\partial v_{GS}^2} + \frac{\partial^2 F}{\partial v_{GS}^2} + \frac{\partial^2 F}{\partial v_{GS} \partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial v_{GS}} = 0 \quad (\text{F.10})$$

from which the second-order derivative is found:

$$\frac{\partial^{2} v_{DSAT}}{\partial v_{GS}^{2}} = \frac{-\frac{\partial^{2} F}{\partial v_{GS}^{2}} - \frac{\partial^{2} F}{\partial v_{DSAT}^{2}} \left(\frac{\partial v_{DSAT}}{\partial v_{GS}}\right)^{2} - \frac{\partial^{2} F}{\partial v_{GS} \partial v_{DSAT}} \cdot \frac{\partial v_{DSAT}}{\partial v_{GS}}}{\frac{\partial F}{\partial v_{DSAT}}}$$
(F.11)

The quantities in the right-hand side are known or they can be easily determined when the expression of F is known.

Other derivatives, for instance derivatives with respect to  $v_{SB}$  or cross-derivatives, can be obtained in a similar way.

The above derivations can be extended easily when velocity saturation is taken into account. In that case, the drain current can be written as a product of three functions that depend on the terminal voltages, as discussed in Section 7.7.2:

$$i_D = mobred(v_{GS}, v_{DS}, v_{SB}) \cdot hot(v_{GS}, v_{DS}, v_{SB}) \cdot large(v_{GS}, v_{DS}, v_{SB})$$
 (F.12)

When for instance the derivative of the current with respect to  $v_{GS}$  is to be known, then, again, this product of three functions can be differentiated with respect to  $v_{GS}$ . The resulting expression will again contain derivatives of  $v_{DSAT}$  with respect to  $v_{GS}$ . The latter derivatives can be evaluated as given in equations (F.9) and (F.11), now of course with a function F that takes into account velocity saturation.

#### **Appendix G**

## Derivation of the MOS drain current in the presence of velocity saturation

In this appendix expressions for the drain current in the triode regime are determined using different models of velocity saturation, which are repeated here for convenience. The model of equation (7.86) is given by

$$v = \frac{\mu_{eff} E_x}{1 + \frac{E_x}{E_c}} \tag{G.1}$$

with  $E_c$  being the critical electric field:

$$E_c = \frac{v_{sat}}{\mu_{eff}} \tag{G.2}$$

The model of equation (7.87) is given by

$$v = \frac{\mu_{eff} E_x}{1 + \frac{E_x}{E_c}}, \quad E_x \le E_c$$

$$= v_{sat}, \quad E_x > E_c$$
(G.3)

where  $E_c$  is now given by

$$E_c = \frac{2v_{sat}}{\mu_{eff}} \tag{G.4}$$

The more accurate model of the velocity-field relationship for electrons is given by (see also equation (7.84))

$$v = \frac{\mu_{eff} E_x}{\left[1 + \left(\frac{E_x}{E_c}\right)^n\right]^{1/n}}$$
 (G.5)

with n=2 for electrons and  $E_c$  again given by equation (G.2).

### G.1 Derivation of the drain current with the simple velocity-field models

We begin by a derivation of the drain current with the most simple models, equations (G.1) and (G.3). For both of these models, the derivation of the drain current in the triode region is the same.

The starting point is the expression of the current in terms of the velocity v(x) at point x in the channel  $(0 \le x \le L)$ , given in equation (7.14):

$$i_D = i_D(x) = W(-Q'_I(x)) v(x)$$
 (G.6)

where v(x) is the velocity of carriers at point x and  $Q_I'$  is the charge per unit area of the inversion layer. This charge is given by equation (7.20). For the sake of simplicity in this derivation, the charge of the depletion layer is assumed to vary linearly along the channel, as given in equation (7.45). In this case,  $Q_I'$  reduces to

$$Q'_{I} = -C'_{ox} \left( v_{GB} - V_{FB} - \phi - v_{CB} - \gamma \sqrt{\phi + v_{SB}} - (a - 1) (v_{CB} - v_{SB}) \right)$$
(G.7)

This velocity can be related to the lateral field  $E_x(x)$  at position x, according to the model of equation (G.1) or (G.3):

$$v(x) = \frac{\mu_{eff} E_x(x)}{1 + \frac{E_x(x)}{E_c}}$$
 (G.8)

with  $E_c$  given by equation (G.2) or (G.4) depending on the model that is used. Substituting equation (G.8) into equation (G.6) and solving for  $E_x(x)$  yields

$$E_x(x) = \frac{i_D}{W\mu_{eff}(-Q_I'(x)) - i_D/E_c}$$
 (G.9)

The electric field is the gradient along the channel of the voltage difference between point x in the channel and the bulk:

$$E_x(x) = \frac{dv_{CB}}{dx} \tag{G.10}$$

Hence, we obtain

$$i_D dx = (W\mu_{eff} (-Q_I'(x)) - i_D/E_c) dv_{CB}$$
 (G.11)

Integrating equation (G.11) from x=0 to x=L and  $v_{CB}=v_{SB}$  to  $v_{CB}=v_{DB}$  and using equation (G.7) yields the final expression for the current in terms of voltages referred to the source

$$i_D = \frac{W}{L} \frac{\mu_{eff} C'_{ox} \left[ (v_{GS} - V_T) v_{DS} - \frac{1}{2} a v_{DS}^2 \right]}{1 + v_{DS} / (LE_c)}$$
(G.12)

When performing the integration,  $\mu_{eff}$  was assumed to be independent of the position in the channel. This also implies that mobility reduction due to the vertical field and velocity saturation are assumed to be independent.

### G.2 Derivation of the drain current with the more accurate velocity-field model

First the same approach will be followed as in the previous section. This will lead to an implicit expression for the drain current which cannot be solved analytically. In order to obtain an analytic expression, a simplified approach will be followed in Section G.2.2.

#### **G.2.1** The rigorous approach

According to the approach followed in Section G.1, the current can be expressed as a function of the lateral electric field, using equations (G.5) and (G.6). This yields

$$i_{D} = \frac{W(-Q'_{I}(x)) \mu_{eff} E_{x}(x)}{\sqrt{1 + \left(\frac{E_{x}(x)}{E_{c}}\right)^{2}}}$$
(G.13)

Squaring this expression and solving for  $(E_x(x))^2$ , one obtains

$$(E_x(x))^2 = \frac{i_D^2}{W^2(Q_I'(x))^2 \mu_{eff} - \frac{i_D^2}{E_x^2}}$$
(G.14)

or, using equation (G.10)

$$\frac{dv_{CB}}{dx} = \frac{i_D}{\sqrt{W^2(Q_I'(x))^2 \mu_{eff} - \frac{i_D^2}{E_c^2}}}$$
(G.15)

From this equation the current can be obtained by integrating from x = 0 to x = L and  $v_{CB} = v_{SB}$  to  $v_{CB} = v_{DB}$ :

$$\int_{0}^{L} i_{D} dx = \int_{v_{SB}}^{v_{DB}} \sqrt{W^{2}(Q'_{I}(x))^{2} \mu_{eff} - \frac{i_{D}^{2}}{E_{c}^{2}}} dv_{CB}$$
 (G.16)

If the charge of the depletion layer is assumed to vary linearly along the channel, then the integral in the right-hand side of equation (G.16) is of the form

$$\int_{p}^{q} \sqrt{ax^2 + bx + c} \ dx$$

This integral can be computed analytically, leading to a complicated expression that contains the terminal voltages of the transistor and  $i_D$  as well. Hence equation (G.16) reduces to

$$i_D = G(v_{GB}, v_{DB}, v_{SB}, i_D)$$
 (G.17)

This is an implicit equation in  $i_D$ , which must be solved iteratively. However, the number of iterations can be kept very low if the start value for  $i_D$  in the iteration is the drain current computed with the expression of the drain current that has been derived with the simpler models for velocity saturation.

The next problem to be solved is the determination of the nonlinearity coefficients with the model for the current. As explained in Chapter 3, these are proportional to the derivatives of the current with respect to the terminal voltages of the transistor. In order to determine the derivatives of the current, the two sides of equation (G.17) are derived with respect to the voltage of interest. Assume we want to know the derivative of the current with respect to  $v_{GB}$ . Differentiating both sides of equation (G.17) with respect to  $v_{GB}$  yields

$$\frac{\partial i_D}{\partial v_{GB}} = \frac{\partial G}{\partial v_{GB}} + \frac{\partial G}{\partial i_D} \cdot \frac{\partial i_D}{\partial v_{GB}}$$
 (G.18)

Solving for  $\partial i_D/\partial v_{GB}$  yields

$$\frac{\partial i_D}{\partial v_{GB}} = \frac{\frac{\partial G}{\partial v_{GB}}}{1 - \frac{\partial G}{\partial i_D}} \tag{G.19}$$

From this equation the higher order derivatives can be obtained by further differentiation.

If in equation (G.16) the more accurate expression for  $Q'_I(x)$  is substituted, then an elliptical integral is obtained which in general cannot be computed analytically.

As a conclusion, it can be stated that the approach followed in this section leads to an accurate solution which, however, is difficult to use. In the next section, an approximate approach will be presented which only introduces small errors.

#### G.2.2 Approximate approach

An approximate closed-form expression for the current can be obtained starting from equation (G.13), which is rewritten in the following form

$$i_D \sqrt{1 + \frac{1}{E_C^2} \left(\frac{dv_{CB}}{dx}\right)^2} dx = W\left(-Q_I'(x)\right) \mu_{eff} dv_{CB}$$
 (G.20)

Integrating both sides of this equation from x = 0 ( $v_{CB} = v_{SB}$ ) to x = L ( $v_{CB} = v_{DB}$ ) yields

$$\int_{0}^{L} i_{D} \sqrt{1 + \frac{1}{E_{C}^{2}} \left(\frac{dv_{CB}}{dx}\right)^{2}} dx = \int_{v_{SB}}^{v_{DB}} W\left(-Q_{I}'(x)\right) \mu_{eff} dv_{CB}$$
 (G.21)

Consider now the left-hand side of this equation. This contains  $dv_{CB}/dx$ , which is the electric field at point x in the channel. According to the mean-value theorem, there exists a mean value

of  $dv_{CB}/dx$ , say  $E_x^*$ , such that the left-hand side is given by

$$\int_{0}^{L} i_{D} \sqrt{1 + \frac{1}{E_{C}^{2}} \left(\frac{dv_{CB}}{dx}\right)^{2}} dx = i_{D} L \sqrt{1 + \frac{1}{E_{C}^{2}} \left(E_{x}^{*}\right)^{2}}$$
 (G.22)

This approximation is also made in [Pat 90].

The right-hand side of equation (G.21) can be easily integrated using either the simplified or accurate expression for the inversion layer charge (equation (7.20) or (G.7)).

Let's now make a simplifying approach by stating that the mean value of the electric field  $E_x^*$  is given by  $v_{DS}/L$ . This assumption would mean that the electric field is constant along the channel. It is the same assumption that is made in the computation of the effective mobility (Section 7.6.2). With this assumption, which is most satisfactory at low  $v_{DS}$  values the current now becomes

$$i_{D} = \frac{\mu_{eff} C'_{ox} W}{L} \frac{1}{\sqrt{1 + \left(\frac{v_{DB} - v_{SB}}{L E_{c}}\right)^{2}}} \left\{ (v_{GB} - V_{FB} - \phi) (v_{DB} - v_{SB}) - \frac{1}{2} \left(v_{DB}^{2} - v_{SB}^{2}\right) - \frac{2}{3} \gamma \left[ (\phi + v_{DB})^{3/2} - (\phi + v_{SB})^{3/2} \right] \right\}$$
(G.23)

#### **Index**

AC beta (BJT), 316	BSIM3, see MOS model, BSIM3
AC conductance, 38	bulk-drain capacitance, 296
AC resistance, 37	bulk-source capacitance, 295
admittance matrix, 120, 123, 162	•
amplifier	capacitance
basic, see basic amplifier	nonlinear, 27, 29, 30, 38–39
operational, see operational amplifier	capacitor
single-transistor	integrated, 52–55
BJT, viii, 119, 303–348, 383, 406, 407	junction, 52
MOS, viii, 350-366, 383, 401, 405,	capacitors
407	in a bipolar transistor, 196
two-stage, 88	in a MOS transistor, 292–296
	extrinsic, 293–294
balanced circuit, 10, 80	intrinsic, 294–296
barrier lowering, 278, 279	cascade connection
base conductivity modulation, 181, 191, 199	of nonlinear circuits, vii, 85–88, 114
base current, 33, 188–190	cascode transistor, viii, 303, 397-401
base pushout, 181, 191	channel
base resistance, 181, 190-196	of a MOS transistor, 210
AC, 193	channel-length modulation, 203, 226, 227,
DC, 193	284–286
extrinsic, 55, 191–193, 319	factor, 202, 213, 226
intrinsic, 55, 191–195, 319	characteristic length, 280
model of Chiu, 193, 195, 199	charge sharing, 278
power model, 192, 193, 199	class-AB operation, 11
base-collector capacitance, 181, 187, 196, 333	CLM, see channel-length modulation
base-emitter capacitance, 181, 196, 325	CMNA, see nodal analysis, compacted mod-
base-width modulation, 191	ified
basic amplifier, 93	CMNA matrix, 121, 123–125, 128, 135, 139,
basic nonlinearities, 27, 29	141, 142, 144, 162, 370
power series description, 29–47	collector current, 32–33, 40–41, 182–188
BJT transistor models, 180–199	collector resistance, 181
body-effect coefficient, 202, 212, 273, 275,	collector-substrate capacitance, 181, 196, 333
276	common-base-transistor, viii, 303, 401-407
box doping profile, 273	common-emitter amplifier, see amplifier, single-
BSIM, see MOS model, BSIM	transistor, BJT

367, 376, 407

common-gate transistor, viii, 303, 401-407	depletion layer, 210, 273
common-source amplifier, see amplifier, single-	charge, 212
transistor, MOS	nonuniform, 217, 228
compacted modified nodal analysis, see nodal	linear approximation, 221-223, 229
analysis, compacted modified	uniform, 212, 226
comparator, 11	describing function, 61
compression, 8, 10, 13, 76	desensitization, 22, 76
compression point, 13, 14, 20	DIBL, see drain-induced barrier lowering
conductance	differential pair
nonlinear, 26, 29–34	BJT, viii, 81–83, 137, 303, 366–375, 377,
conversion	383, 385
between voltage-controlled and current-	MOS, viii, 303, 375–386
controlled description, 35	diffusion capacitance, 39, 181, 197, 198
conversion gain, 24, 146, 415, 424	direct calculation of nonlinear responses, 137-
cross modulation, 24, 60, 79	156, 177
factor, 25	.DISTO, 164
amplitude modulation, 79	distortion, 4, 8
phase modulation, 79	harmonic, see harmonic distortion
crossover distortion, 6	linear, 4
current crowding, 181, 191, 199	nonlinear, 4, 8
current divider	doping
nonlinear capacitive, 118, 171–175	box profile, see box doping profile
current mirror, 91, 92, 303, 407–414	gaussian profile, see gaussian doping pro-
BJT, viii, 407–409, 413–414	file
MOS, viii, 408–413	nonuniform, see nonuniform doping ef-
current source	fects
nonlinear	downconversion, 23
of order three, <i>see</i> current source, non-	downconverter, 81, 84, 147
linear, third-order	drain current, 42–45, 203–292, 297–300
of order two, see current source, non-	drain-induced barrier lowering, 203, 278, 284,
linear, second-order	285, 287
nth-order, 134	drift velocity, 210, 225, 245, 247, 253
second-order, 122, 124, 132, 135, 143–	dynamic range, 13
145, 156	intermodulation-free, see intermodulation
third-order, 126, 127, 132, 133, 135–	free dynamic range
137, 147–149, 157	noe dynamie range
cutoff frequency, 333	Early effect, 181, 182, 187, 310, 355
cutoff frequency, 555	nonlinear, 184, 199, 436, 478
DC conductance, 38	Early resistance, see Early effect
DC resistance, 37	Early voltage, 182, 187, 285-287, 289
DC shift, 8, 76	electric field
DC transfer characteristic, 82, 153, 154, 304,	critical, 247

lateral, 210

longitudinal, see electric field, lateral normal, 231 vertical, see electric field, normal emission coefficient, 182, 189, 306 emitter crowding, see current crowding emitter degeneration, viii, 100–103, 390 nonlinear, 395 emitter follower, viii, 162, 177, 386–390, 392, 394, 397 emitter resistance, 181 even-order distortion, 16 suppression, see suppression, of even-order distortion expansion, 8, 13, 76	harmonic, v second, 8, 76 third, 8, 76 harmonic balance method, 175–178 harmonic distortion, 8, 77 second-order, 8 third-order, 8 total, 14 harmonic suppression, 240 high injection, 182–184, 189, 190, 306, 321, 324, 325 hot function, 249, 250 for BSIM model, 250 merging with mobred, 250 hot carriers, 245, 282
feedback, 2, 93–114, 386, 428 active, 114 linear, vii, 97–103, 105 nonlinear, vii, 113–114 positive, 98 with a large loop gain, 96–97 feedback network, 93 Fermi potential, 211 flat-band voltage, 211 forward transit time, 198 Fourier transform multidimensional, 66, 469 p-dimensional, 468 two-dimensional, 463 front-end analog, iv, v, 3, 22, 23, 177 receiver, 3, 22, 23	image frequency, 4 impact ionization, 281, 285 intercept point input, 13 output, 13 second-order for harmonics, 13, 79 for intermodulation products, 21 third-order for harmonics, 13, 79 for intermodulation products, 21, 79 intermodulation distortion, 18, 77 second-order, 18 third-order, 18 intermodulation product, v second-order, 17, 76
gain reduction factor (feedback), 94 gate-bulk capacitance, 296 gate-drain capacitance, 296 gate-oxide thickness, 202, 212 gate-source capacitance, 295, 409 gaussian doping profile, 273, 278 Gilbert multiplier, 84–85, 114, 414 Gummel-Poon model, viii, 180, 181, 183, 199, 304, 478	third-order, 17, 76 intermodulation-free dynamic range, 22 inverse system, 89, 114 inversion layer, 210, 273

kernel transform	MNA, see nodal analysis, modified
nth-order, 65	MNA matrix, 121, 123, 125, 135, 139, 144,
pth-order, 468	162, 370
second-order, 463	mobhot function, 256
knee current, 184, 321	mobility, 245
Kronecker product, 117, 137	effective, 231–245, 249, 251
Krylov subspace method, 177	model of Frohman-Bentchkowsky, 236
Krylov suospace method, 177	model of Liang, 237
Laplace transform	reduction, 231–245, 257
multidimensional, 65, 469	critical field, 202, 237
p-dimensional, 468	surface, 202, 210, 211
two-dimensional, 463, 464	mobility reduction, 277
large function, 249, 250	coefficient, 202, 232
for BSIM model, 250	mobred function, 234, 249, 250
late effect, 478	for BSIM model, 250
LDD, see lightly-doped drain	merging with hot, 250
level 1, see MOS model, level 1	moderate inversion operation, 201
level 2, see MOS model, level 1	modified nodal analysis, see nodal analysis,
level 3, see MOS model, level 1	modified
lightly-doped drain, 200, 280–282, 284, 287	modulation index, 24
linear distortion, see distortion, linear	MOS model
loading effect	BSIM, 202, 221, 222, 226, 229, 237, 250,
of a nonlinear feedback network, 103–	276, 300
105	BSIM3, 203, 250, 280, 284, 287–289,
low-distortion conditions, 11	355
low-distortion conditions, 11	BSIM3 version 2, 289, 290, 292
matching, 160, 373, 385, 414, 417	BSIM3 version 3, 289
memory	Frohman-Bentchkowsky, see MOS model
circuit with, 10	level 2
memoryless circuit, 10, 69, 70, 75, 77, 78,	level 1, 200, 202, 203, 213, 223, 224,
88, 90, 97, 98	227, 229, 230, 244, 257, 258, 264,
Miller	269, 284, 300, 352, 354, 356, 376,
compensation, 106, 159, 303, 312, 356,	382, 383
417	level 2, 202, 203, 221, 223, 224, 226
effect, 336, 347	228–230, 233, 236, 237, 279, 284,
factor, 335, 347	300
mismatches, 154, 303, 367, 372, 375, 384,	level 3, 202, 203, 221–224, 226, 229
385, 407, 409, 411, 412, 422	230, 235, 238, 241, 243, 279, 284
mixer, 23, 115, 137, 154	300
double-balanced, viii, 80, 81, 303, 414	Shichman and Hodges, see MOS model
single MOS transistor, 216, 420–427	level 1
switching, 23	MOS transistor models, 200-301
weakly nonlinear, 23, 216, 414, 420	multidimensional

Fourier transform, see Fourier transform, multidimensional	numerical integration, 175 numerical verification
Laplace transform, see Laplace transform, multidimensional	of steady-state nonlinear behavior, 175
multidimensional transconductance, see transcon-	odd-order
ductance, multidimensional	distortion, 16
multiple-input system, 117, 137, 179, 459	suppression, see suppression, of odd- order distortion
nodal analysis, 120	offset voltage, 374
compacted modified, 121	opamp, see operational amplifier
modified, 121	operational amplifier, 4, 15, 105-113, 159,
nonlinear	303, 417–419, 431, 434
capacitance, see capacitance, nonlinear	fully-differential, 159, 160
conductance, see conductance, nonlin-	operator
ear	bilinear, 460, 461
transconductance, see transconductance,	p-linear, 467
nonlinear	second-order, 460
transfer function, see transfer function,	trilinear, 467
nonlinear	Volterra, see Volterra operator
nonlinear behavior	output conductance
second-order, 8	of a BJT, 186, 187, 311
strongly, v, 6, 11	of a MOS transistor, 215, 226, 283-292
third-order, 8	
weakly, 11, 67	perturbation method, 117
nonlinearity coefficient, vii	poly-gate depletion effect, 203
first-order, 31, 35, 38	post-distortion, vii, 88–92, 115, 318, 409
normalized, 32, 46, 51, 55, 182, 184, 187, 196	pre-distortion, vii, 88–92, 115, 318, 409, 411, 414, 428
second-order, 32, 35, 38, 147, 215, 227, 228, 239, 266, 269, 293, 299, 309,	quasi-saturation, 180
351, 353	relative error
third-order, 32, 35, 38, 215, 228, 266,	signed, 223, 259, 269–272
270, 294, 299, 309, 351, 353	resistance
of a circuit	base, see base resistance
higher-order, 9	collector, see collector resistance
second-order, 9	emitter, see emitter resistance
third-order, 9	nonlinear, 26, 29, 30, 34–38
second-order, 31, 35, 38	of source and drain regions, 203
third-order, 31, 35, 38	source and drain, 280–283
nonuniform doping effects, 203, 266, 273-	resistor
278	integrated, 47–52
nonuniform substrate, see nonuniform dop-	diffused, 48–52
ing effects	implanted, 48-52

saturation current	tensor, 117
base-emitter leakage, 189	third-order signal, 10
saturation current (BJT), 33, 39, 182, 186,	third-order system
375, 408	basic, 69
saturation region (MOS), 225	sinusoidal response, 469
saturation velocity, 202, 245	threshold voltage, 202, 213, 273, 352, 385,
saturation voltage (MOS), 225	411
SC, see substrate current	models for nonuniform substrates, 275
second-order signal, 10	piecewise model, 275
second-order system	roll-off, 278–280
basic, 68	short and narrow devices, 278–280
sinusoidal response, 464, 466	zero-bias gate-source extrapolated, 202
shooting method, 175, 176	tracking nonlinearities, 45–47, 167, 173, 175,
simplification	190, 198, 317, 319, 327, 400, 401,
after generation, 160	414
before generation, 160	transcapacitance, 30
during generation, 160	nonlinear, 27
slew rate, 6, 11	transconductance
source follower, viii, 391–397	multidimensional, 26, 29, 30
spreading resistance effect, 281	nonlinear, 26, 29–34
spurious responses, 18	three-dimensional, 41–45
square waveform, 14	two-dimensional, 39–41
square-law model, 227	transfer function
strong inversion operation, 201	nonlinear, 65
strongly nonlinear behavior, see nonlinear be-	nth-order, 65, 66
havior, strongly	transistor models
substrate current, 203, 284, 285, 288	weakly nonlinear, 55
subthreshold region, see weak inversion op-	transit time, 39, 198
eration	transresistance
suppression	multidimensional, 26
of even-order distortion, vii, 2, 16, 80–	nonlinear, 26
85, 114, 428	triangular waveform, 15
of odd-order distortion, vii, 2, 16, 80–	triode region (MOS), 210
85, 114	thode legion (WOS), 210
surface inversion potential, 202, 211	upconversion, 23
surface mobility, 202	upconverter, 84, 146, 148, 416, 420
surface potential, 297	CMOS viji 420–427
symbolic analysis, 117	CMO3, VIII, 420–427
symbolic computation	VBIC95 model, 180
of nonlinear responses, 156–164	velocity saturation, 245–272
symbolic network analysis, 158	velocity-field model
symmetrization of Volterra kernels, see Volterra	accurate model, see velocity-field mode
kernel, symmetrization	model 1

```
model 1, 247, 251, 256
   model 2, 247, 249, 252, 254-256
    model 3, 248, 249, 252, 255
    piecewise model, see velocity-field model,
        model 3
velocity-field models, 247-248
voltage coefficient, 47, 55, 187
voltage divider
    nonlinear, 118, 164-171, 325, 332
Volterra
    kernel, 60, 68, 459
      nth-order, 64
      calculation, 119-137
      even-order, 80
      odd-order, 80
      of a current, 132
      pth-order, 468
      second-order, 63, 462
      symmetric, 64, 66
      symmetrization, 463, 468
    operator, 62-63, 459
      first-order, 63
       nth-order, 64
       p-th order, 467
       second-order, 63, 460
       third-order, 63
     series, vi, vii, 1, 2, 7, 11, 20, 59-115,
         458, 459
 weak inversion operation, 201, 203, 297
 weak inversion slope, 298
 weakly nonlinear
     mixer, see mixer, weakly nonlinear
 weakly nonlinear behavior, see nonlinear be-
         havior, weakly
 weakly nonlinear circuit, 11
```